1-1-2006

Detecting exposed items in computer-based testing.

Ning Han
University of Massachusetts Amherst

DETECTING EXPOSED ITEMS IN COMPUTER-BASED TESTING

A Dissertation Presented

by

NING HAN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 2006

Education

DETECTING EXPOSED ITEMS IN COMPUTER-BASED TESTING

A Dissertation Presented

by

NING HAN

Approved as to style and content by:

Ronald K. Hambleton, Chair

Lisa A. Keller, Member

Hui-Kuang Hsieh, Member

Christine B. McCormick, Dean
School of Education

DETECTING EXPOSED ITEMS IN COMPUTER-BASED TESTING

A Dissertation Presented

by

NING HAN

Approved as to style and content by:

Ronald K. Hambleton, Chair

Lisa A. Keller, Member

Hui-Kuang Hsieh, Member

Christine B. McCormick, Dean
School of Education

DEDICATION

To my dear wife, Hui, for her love and selflessness. It would have been impossible for me to stay in this remote country without her presence and support. To my parents, I extend my gratitude, for their endless encouragement and patience.

ACKNOWLEDGMENTS

The completion of this dissertation is an important personal accomplishment for me. It was a hard decision to go back to the classroom after I had left school for more than 14 years. Luckily, I made a correct decision.

I begin my acknowledgements by expressing my deepest appreciation to the chair of my dissertation committee, my adviser throughout graduate school, and my mentor in multiple dimensions, Professor Ronald K. Hambleton. He was my primary reason to come to UMASS several years ago, and he has given me so much, not only considerable knowledge and skills, but also a philosophy and attitude that goes far beyond psychometrics. This study, together with many other projects that Ron has involved me in, has benefited considerably from his wisdom.

As a member of my committee and a REMP professor, Lisa Keller has always provided me with exceptional guidance and the great flexibility to finish this dissertation. I felt guilty every time I saw her since I am far below her expectations. My work at ETS was valuable, but also the reason for finishing the thesis later than she expected.

I also thank Professor Hui-Kuang Hsieh for the significant insights he has contributed to the study. To another very important faculty in REMP: Professor Stephen Sireci for his enthusiasm.

It was painful when I knew Swami was retiring and leaving our program. I was very lucky to have the opportunity to take his classes but I was not lucky enough to have his insight for my last two years at UMASS.

I am grateful to all the students who shared Hills South with me: Dean Goodman, April Zenisky, Billy Skorupski, Shameem Khaliq, Ying Lu, and Shuhong Li. Many of you shared of yourselves and your lives making my life in this small town cheerful.

Finally, I want to thank two of the biggest testing organizations in the world that I have worked for: National Education Examinations Authority (NEEA) of China and Educational Testing Service (ETS) of the USA, and I am grateful to both the AICPA and the Massachusetts Department of Education for financially supporting my UMASS research. All of this is impossible without my wonderful working experiences with my mentors, friends, and colleagues in these organizations.

ABSTRACT

DETECTING EXPOSED ITEMS IN COMPUTER-BASED TESTING

MAY 2006

NING HAN, B. S., EAST CHINA NORMAL UNIVERSITY

Ed. D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

More and more testing programs are transferring from traditional paper and pencil to computer-based administrations. Common practice in computer-based testing is that test items are utilized repeatedly in a short time period to support large volumes of examinees, which makes disclosed items a concern to the validity and fairness of test scores. Most current research is focused on controlling item exposure rates, which minimizes the probability that some items are over used, but there is no common understanding about issues such as how long an item pool should be used, what the pool size should be, and what exposure rates are acceptable.

A different approach to addressing overexposure of test items is to focus on generation and investigation of item statistics that reveal whether test items are known to examinees prior to their seeing the tests. A method was proposed in this study to detect disclosed items by monitoring the moving averages of some common item statistics.

Three simulation studies were conducted to investigate and evaluate the usefulness of the method. The statistics investigated included classical item difficulty, IRT-based item raw residuals, and three kinds of IRT-based standardized item residuals.

The detection statistic used in study 1 was the classical item difficulty statistic. Study 2 investigated classical item difficulty, IRT-based item residuals and the best known of the IRT-based standardized residuals. Study 3 investigated three different types of standardizations of residuals. Other variables in the simulations included window sizes, item characteristics, ability distributions, and the extent of item disclosure. Empirical type I error and power of the method were computed for different situations. The results showed that, with reasonable window sizes (about 200 examinees), the IRT-based statistics under a wide variety of conditions produced the most promising results and seem ready for immediate implementation. Difficult and discriminating items were the easiest to spot when they had been exposed and it is the most discriminating items that contribute most to proficiency estimation with multi-parameter IRT models. Therefore, early detection of these items is especially important. The applicability of the approach to large scale testing programs was also addressed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Since the mid-1990s, dramatic reductions in the sizes and costs of computers have increased their availability and led to their application in most aspects of our lives. During this same period, advances in educational and psychological assessment have led to a revolutionary innovation, computer based testing (CBT) (e.g., Mills, Potenza, Fremer, & Ward, 2002).

The computer has been applied to educational and psychological measurement since its beginning as a powerful data processing tool. Along with the development of more complicated psychological models and theories, it becomes helpful to increase the statistical accuracy of test scores. The emergence of so-called CBT not only refers to the applications of the computer technology as a data process tool or administrative assistant, but also initiates a new era in the field of educational and psychological measurement.

CBT can probably bring a lot of advantages, which were unimaginable in paper and pencil tests (Weiss, 1983; Drasgow & Olson-Buchanan, 1999; Mills, Potenza, Fremer, & Ward, 2002). Currently, more and more education and psychology assessment programs are being converted from traditional paper and pencil based tests into computer-based tests (e.g. Graduate Record Exam, Test of English as a Foreign Language).

Computers, as perhaps one of the most pervasive of the present technologies, can contribute to assessments in many ways. Computer programming affords test developers the flexibility of dynamic selection of items to be presented and allows variations in the presentation of stimulus materials. They can present single items for a limited period of time (to limit exposure) or tailor the exam to the examinee's ability (i.e. adaptive testing) (Drasgow & Olson-Buchanan, 1999). They can also include simulations of real-life situations, graphics and videos, and voice-activated responses so as to create innovative item formats and more realistic testing environments.

Although CBT offers many advantages over traditional paper and pencil tests, new concerns about test security have evolved. Generally speaking, unlike the paper and pencil test, which is usually administered in a limited time period, CBT is administered in much wider time windows to support the large volume of examinees. It becomes critical to assure that items in an item pool are safeguarded from being disclosed to the examinees before the examinees take the test. If this problem is not addressed properly, it may dramatically reduce the usefulness of CBT in high stakes testing.

Traditionally, paper and pencil tests have maintained test security primarily by "lock and key". New test forms are regularly developed for new administrations and each administration is exposed for only a relatively short period of time. The test items have very little possibility to be exposed prior to the date of administration. However, since test items have to be utilized daily in order to support a large volume of test takers, this type of controlled access is no longer working in CBT. It is impractical to develop unique test forms for each administration. On the contrary, the same items have to be

administered over multiple days, weeks, even months. This greatly enhances the possibility of exposing the items and obviously is one of the biggest threats to the validity and fairness of CBT.

The standards of most professions that use test materials require that secure tests not be disclosed in a forum where they are accessible to individuals who may be taking the test at some point in the future. For example, Standard 5.7 of the recently revised Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) states, "test users have the responsibility of protecting the security of the test materials at all times" (see also Standard 11.7). Furthermore, Standard 5.6 states "reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means." The psychometric integrity of these instruments depends upon the test taker having no prior access to the test questions and answers.

In spite of the potentially serious consequences, test developers have very few tools to address the challenge. In traditional non-adaptive test designs, all examinees see the same or an equivalent test from that consists of a series of items that measure the construct of interest. The test administrator faces the same situation with the traditional paper and pencil test but it is more difficult to guard the test items because of the cost of test administration. This is the primary reason that the traditional linear test design is rarely seen with CBTs. More complicated test designs such as computer-adaptive test (CAT) or multi-stage test (MST) designs were introduced to gain the advantage of CBT.

In these test designs, different examinees usually are administered different test items. The number of items to achieve the required accuracy can be reduced dramatically if only those items providing the most statistical information in proficiency estimation for particular examinees are used. (Of course, content validity is still a factor too in item selection.) However, not every item in an item pool provides the best statistical information. The optimal item selection principle will yield an absurd result in that some items will be administered frequently while some other items will be seldom used.

A widely used operational technique, item exposure control, then was introduced to address the problem of overexposing some items. The primary reason to apply item exposure control is to address the concern about test item security.

Some other research studies can be found that address the issue of item security. The drift of item parameters not only reflects the change of trend of examinee abilities but also may result from exposure of test items. An examinee who exhibits inconsistent response patterns may know correct answers to some items before the test administration. However, most of the research studies have overlooked the possibility of item exposure as a main explanation for the finding. Clearly, more research about the detection of exposed test items would seem to be in order.

1.2 Purposes and Significance of the Study

There are two facets to the problem of test security. One is how to prevent items in item pools from theft. Almost all the research that has been done can be classified into this category. The other direction is how to spot an item once it is known by examinees. These two facets are both important. Han (2003) proposed to detect exposed

test items using the "moving averages" of some item statistics in an earlier study for the American Institute of Certified Public Accountants.

In this method, item performance can be monitored over time (e.g., after each item administration), and any changes can be noted and used to identify potentially exposed test items. Preliminary research has been encouraging. With a moderate examinee sample size, an exposed item can be spotted in a relative short time window. At the same time, this research has been based upon the assumption that the ability distribution of the examinees who take the CBT over time is stationary (Han, 2003), which is hardly met in practice. The current research will expand the work of Han (2003) in several directions: investigating additional item exposure statistics, and evaluating these statistics under different conditions such as with shifting ability distributions over time and with various types of items (e.g., hard or easy, with low or high discrimination), and for several exposure models.

More specifically, the purposes of this research were: (1) to propose and adapt several item exposure detection statistics including classical test theory based and item response theory (IRT) based. (2) to investigate and evaluate the performances and properties of the detection statistics with different ability distributions including fixed ability distributions and the presence of shifts in the ability distribution over time, (3) to address the suitability of the item exposure detection statistics under a number of item exposure models, and (4) to investigate item exposure detection for items with different statistical characteristics.

These purposes are essential because the assumption to assume a fixed ability distribution at all times during a testing window made in Han (2003) is too strong for

most education testing programs. Some drift in the distribution might be expected—for example, the poorer candidates may test first, and higher ability candidates may follow later in the window. Several new item exposure statistics needed to be proposed and investigated because the moving p-value statistic that Han (2003) considered was sensitive to ability shifts, and therefore, it was less suitable for use in practice. Achieving the second and third purpose would provide data on competing item exposure detection statistics under various item exposure models. For example, in one simple model, after an item is exposed by a candidate one might conjecture that all candidates will have knowledge of the item and answer it correctly if it is selected for administration again. Several other more realistic item exposure models needed to be investigated too. The fourth purpose was added to the study because it was expected that the item exposure detection rate would depend not only on the choice of item exposure detection statistic, sample size, and nature of the exposure, but would also depend on the statistical characteristics of the exposed test items. For example, it was expected that it would be very difficult to detect exposed items when they were easy for candidates. After all, candidates are already expected to do well, and any improvements in item performance due to exposure then would be small, and harder items should be considerably easier to spot because the shifts in item performance due to exposure are likely to be greater.

This method of "moving averages" focused on the other facet of test security to spot an item once it has been disclosed to examinees. It cannot prevent test items from being compromised, but it will set up an alarm system to monitor the performance of

test items so that test developers and administrators can be aware of where they are, which enables them to take the according suitable reactions on time.

CHAPTER 2

REVIEW OF THE LITERATURE

2.1 Introduction

The computer has being used in the administration of educational and psychological tests for decades like in other fields. Nevertheless, CBT does not imply that the computer is used as a data processing tool for administrative purposes but as a test delivery mechanism. The necessity of CBT can be traced to Binet's major advance in intelligence testing. Since his concern was with the diagnosis of the individual candidate, he realized that he could tailor the test to the individual by a simple stratagem - rank ordering the items in terms of difficulty (Binet, 1905). Lord's (1980) work on computer-based testing is a refinement of Binet's method and Lord's procedure can be conveniently operated by personal administration or by a computer. The items are stratified by difficulty level, and several subsets of items are formed at each level. The test then proceeds by administering subsets of items, and moving up or down in accord with examinee success rate on each subset. After the administration of several subsets, the final candidate ability estimate is obtained. However, in Lord's era, the scarcity, expense, and awkwardness of computer hardware and software limited the implementation of CBT. But this situation was changed by the middle 1990s.

Educational Testing Service (ETS) is perhaps the first educational testing agency in the world trying to convert large-scale high stakes testing into CBT. From the beginning of the conversion, test security has been an important topic but early theoretical investigations of CAT ignored the problem of test security (see, for example, Lord, 1970), until an incident happened in the early stage of operational CBT, which

demonstrated how the security problem could seriously affect the validity and fairness of CBT. Davey and Nering (2002) reported this incident: After sending its employees to take the GRE tests and memorize as many items as possible for a short period of time, Kaplan Educational Center discovered that most of the items its employees collected were already on the list of compromised items. ETS had to shut down testing temporarily after Kaplan notified ETS that a large portion of the item pool was known to Kaplan. The following is a detailed description, retrieved from the Internet.

> In response to criticism that the computerized version of the Graduate Record Examination (GRE) is easy to cheat on, the Educational Testing Service (ETS), responsible for the test, has reduced the amount of times the tests will be offered via computer. The security problem became an issue when 20 investigators from Kaplan Educational Centers, a test preparation firm, took the exam and were able to reconstruct "a significant portion" of the exam. Jose Ferreira, director of G.R.E. programs at Kaplan, said that the solution depends on whether ETS is willing to spend money to make improvements, such as adding four to five thousand questions, instead of just recycling questions that can be memorized and passed along. Nancy S. Cole, president of ETS, countered by accusing Kaplan of having a vested interest in exposing computer-testing flaws. She said that students who take computerized tests "tend to prepare on their own rather than en masse." ETS filed suit to keep Kaplan from sending investigators. Kaplan agreed to keep investigators out while its officials meet with the testing service to try to settle the lawsuit. (CHANCE News 4.01)

Clearly, the major security weakness of CATs lies with continuous testing. Indeed, since the Kaplan-ETS incident, people have started to realize how vulnerable CATs could be to organized attempts to memorize items when testing is ongoing. The ETS-Kaplan incident trigged studies on test security of CAT and certain theoretical justifications have been developed. One natural option the testing company can make is to increase the size of item pools. Stocking (1994) suggested that the item pool size should be 12 times the CAT exam length which Way (1998) referred as a rule of thumb. One the other hand, items in the item pool should be used more sufficiently. Wainer

(2000) pointed out that when every item has an equal probability to be administrated to examinees, test security will reach the maximum. However, Wainer (2000) investigated the item usage in GRE- CAT and found that as few as 12% of items could account for as much as 50% of the functional pool. This explains why the focus of the testing industry and research has been put on a major research direction: item exposure control.

2.2 Item Exposure Control

The hardcore of CBT is the item selection process from an item pool for each examinee during test administration. When CBT was still in the theoretical stage, a universally accepted item selection rule was to select items within the framework of item response theory (IRT) according to the item information function. Once an initial estimate of the ability of the examinee is obtained, the next item is the one, which provides the biggest information around the estimated ability point (Lord, 1980; Hambleton, Swaminathan, & Rogers, 1991). The above ETS-Kaplan law case declared the bankruptcy of this well-known and popular item selection algorithm. If this algorithm is applied in practical testing programs, only a small portion of items are utilized frequently while a large part of the items in the item pool remain idle. Economically, it is not acceptable to have many items idle. For one, exposure rates of items being selected will be very high. More seriously, the ETS-Kaplan case showed that this type of tests can be compromised very easily by some professional thieves. Schnipke and Scrams (1999) simulated a situation in which an organized group of "thieves" took the test, memorized the items they received, and distributed the items to future test takers. Results showed that when regular thieves provided the stolen items, all but the highest ability test takers received inflated ability estimates. When

professional thieves provided stolen items, some low-ability test takers were helped tremendously and received ability estimates at the top of the ability range. All test takers who had relatively high abilities received a substantial number of stolen items and also received ability estimates at the top of the ability range when professional thieves provided the stolen items.

This fact triggered additional research on item exposure control. The item selection algorithm based on the item information function is statistically optimal, but, under this algorithm, a small number of items with good statistics will be selected more frequently than those with moderate statistics. The basic idea of item exposure control is to limit the usage of items with good statistics and increase the usage of items with moderate statistics. Therefore, controlling item exposure rates was exactly stimulated by concerns about test security. Stocking (1993) pointed out that for CAT to be a serious competitor to traditional paper and pencil testing, methods must be developed to limit the exposure of items in order to ensure the fairness to all examinees. To date, quite a few methods have been proposed to control item exposure rate (Davey & Parshall, 1995; McBride & Martin, 1983; Stocking & Lewis, 1995a, 1995b; Sympson & Hetter, 1985). The logic of all these methods is to "randomize" the items to some extent instead of "optimizing," though different statistical models were employed.

McBride and Martin (1983) is one of the earliest studies attempting to control item exposure rate. They developed a very simple 5-4-3-2-1 algorithm to prevent some items from over exposure. For an examinee with an ability estimate, the first item will be chosen randomly from the best five items. The second item will be chosen randomly from the best four items and so on.

Sympson and Hetter (1985) and Hetter and Sympson (1995) tackled the issue of controlling item exposure directly in a probabilistic fashion based on the behavior of items over repeated simulations of a test design with a sample drawn from a typical distribution of abilities and then set up an exposure control parameter for each item. If this item is selected according to the optimal rule whether or not the item is actually administered depends on the exposure control parameter. If this item tends to be administered very often, then the exposure control parameter can be set low, meaning that the item will be less likely to be administered. On the other hand, if this item tends to be rarely administered, the exposure control parameter can be set very high so that the item will more likely to be administered. This procedure considered a test taker randomly sampled from a typical group of test takers and distinguishes between the probability P(S) than an item is selected as the best next item to administer from an ordered list formed by a CAT item selection algorithm, and P(A|S), the probability that an item is administered given than it have been selected. The procedure seeks to control P(A and S) = P(A|S) * P(S) and to insure that the maximum value of P(A) for all items in the pool is less than some value *r*. This r is the desired maximum rate of item usage. This procedure is very important in the development of item exposure control methods since most of other methods are based on it and made some modifications to it.

Davey and Parshall (1995) also set up a parameter for each item but this parameter depends on the usage of all other items. Unlike Sympson and Hetter (1985) and Hetter and Sympson (1995), Davey and Parshall (1995) not only reduced the probability that an item would be overused but also reduced the probability of pairs of items appearing together. This method can reduce the extent to which two tests overlap.

Stocking and Lewis (1998a) remodeled the Sympson and Hetter approach. They followed the Sympson and Hetter approach to set up an exposure control parameter for an item and then used a multinomial model to select the next item to be administered. They proposed exploiting this more robust model to develop exposure control parameters conditional on ability level. The basic model considers, at each phase of testing, the list of items ordered from the most desirable to the least desirable and the associated P(A|S), one of each item in the list. As in the Sympson and Hetter procedure, the same adjustment simulations are required to obtain the exposure control parameters, the P(A|S)s.

Stocking and Lewis (1998b) proposed a conditional multinomial method to control the exposure rate for the examinees with similar ability levels. This method derived the exposure control parameter based on a typical ability level instead of to the whole range of proficiency.

All of the exposure control procedures have potentials in terms of maintaining test security. Some extent of randomizations seems essential otherwise examinees may memorize some special paths to obtain a high grade. New methods on item exposure rate control are continuously being proposed.

In addition to item exposure control, some other research was also found to minimize item usage. For example, expanding the number of test items in a bank (either by hiring extra item writers and/or using item generation rules and algorithms) (see Pitoniak, 2002), rotating item banks, expanded initiatives to reduce sharing of test items on the internet, shortening test administration windows, modifying the test design (with the intent of reducing the number of items that candidates are administered, without loss

of precision (see Zenisky & Hambleton, 2004), better item bank utilization (see Yi & Chang, 2003, on item bank usage), and so on.

The second incident around ETS happened more recently. On August 6, 2002, ETS announced it would temporarily suspend its GRE CAT:

> ETS has temporarily suspended the computer-based GRE and reintroduced paper-based versions into part of Asia including China mainland, Hong Kong, Taiwan, Korea, and India. The decision was based on an investigation that uncovered a number of Asian-language web sites offering questions from live versions of the computer-based GRE. The web sites included both questions and answers illegally obtained by test takers who memorize and reconstruct questions and share them with other test takers. (Paper based GRE, 2002)

Clearly, test security has to be studied in a much broader context. The new emphasis should be whether or not a testing company can identify the problematic items if thieves or unintended examinees result in them being exposed.

2.3 Person Fit

The use of person-fit indices provides another way to detect if the performances of examinees deviate from an underlying IRT model (Meijer & Sijtsma, 1995; Nering & Meijer, 1998). Several statistics have been proposed to investigate the fit of an item score pattern to an IRT model.

Most person-fit statistics are designed to investigate the probability of an item score pattern under the null hypothesis of fitting response behavior with the following general form:

$$\sum_{i=1}^{n} X_i w_i(\theta) - w_0(\theta)$$

where X_i is the binary response to item *i*,

$w_i(\theta)$ and $w_0(\theta)$ are suitable weights (see Snijders, 2000).

Levine and Rubin (1979) proposed a statistic using a form of the log-likelihood function, which was further developed by Drasgow, Levine, and Williams (1985).

$$l = \sum_{i=1}^{n} \left\{ X_i \ln P_i(\theta) + [1 - X_i] \ln \left[1 - P_i(\theta) \right] \right\}$$

Drasgow, Levine, and Williams (1985) proposed a standardized version of l, l_z.

$$l_Z = \frac{l - E(l)}{\sqrt{Var(l)}}$$

where

$$l = \sum \{ X_i \ln P_i(\theta) + [1 - X_i] \ln [1 - P_i(\theta)] \}$$

$$E(l) = \sum \{ P_i(\theta) \ln P_i(\theta) + [1 - P_i(\theta)] \ln [1 - P_i(\theta)] \}$$

$$Var(l) = \sum P_i(\theta)[1 - P_i(\theta)][\ln \frac{P_i(\theta)}{1 - P_i(\theta)}]^2$$

van Krimpen-Stoop and Meijer (2000) proposed several item statistics to assess person fit in CAT. The eight item statistics defined in van Krimpen-Stoop and Meijer (2000) are:

$$T_k^1 = \frac{1}{n} \{ X_{ik} - P_{ik}(\hat{\theta}_{k-1}) \}$$

$$T_k^2 = T_k^1 \{ P_{ik}(\hat{\theta}_{k-1})[1 - p_{ik}(\hat{\theta}_{k-1})] \}^{-\frac{1}{2}}$$

$$T_K^3 = T_k^1 \{ I(\hat{\theta}_{k-1}) \}^{-\frac{1}{2}}$$

$$T_k^4 = \sqrt{k} T_k^1$$

$$T_k^5 = \frac{1}{n} \{ X_{ik} - P_{ik}(\hat{\theta}_n) \}$$

$$T_k^6 = T_k^5 \{ P_{ik}(\hat{\theta}_n)[1 - p_{ik}(\hat{\theta}_n)] \}^{-\frac{1}{2}}$$

$$T_K^7 = T_k^5 \{I(\hat{\theta}_n)\}^{-\frac{1}{2}}$$

$$T_k^8 = \sqrt{k} T_k^5$$

where k stands for the *k*th item in the CAT;

$X_{ik} = 1 \text{ or } 0$ is the binary score of examinee *i* on item *k*;

$P_{ik}(\theta)$ is the probability computed from the hypothesized IRT model;

$I(\theta)$ is the test information function

$\hat{\theta}_n$ is the ability estimate after the examinee takes a test consisting of *n* items.

Among these statistics, we can easily find that statistic 1 and 4 are equivalent with item residuals introduced in Hambleton, Swaminathan, & Rogers (1991) (page 60-63) Statistics 2, 3, 6, and 7 are different forms of standardization of item residuals.

2.4 Item Parameter Drift

Another close research direction to detecting exposed items is item parameter drift though so far most reported research has focused on changes in curricula or examinee populations as explanations. Normally CAT administrations require a large supply of items with accurately estimated statistics in order to sustain its continuous testing. However, a pre-estimated IRT model, which is normally obtained during the process of pretest data analysis, doesn't always correctly capture what underlies a new set of examinee responses to the item. This so called item parameter drift could be caused by many reasons, such as not perfect initial pretest calibration due to estimation methodology or limited calibration sample size, differences in motivation of the test takers between the pretest and the operational test, changes in examinees' learning

experience, and so on. Item compromise is also a possible reason, especially when the item has been administered in a CBT environment. When an item is exposed, obviously it becomes easier for all examinees with prior knowledge of it, and correspondingly, item difficulty will go to one (classical testing theory) and minus infinity (IRT) and the discrimination index will go to zero (both classical testing theory and IRT).

In the past decade, the concern about the negative impact of item parameter drift has led to the development of statistical procedures and indexes to measure the extent of item parameter drift by computing the area between a previously estimated item response function and the corresponding newly estimated item response function. Most of the statistics that have been proposed for monitoring item performance and identifying item drift take either of two basic approaches. The first is to work continuously and cumulatively, analyzing on successive occasions all data collected to date. One example of this approach is to periodically recalibrate an item based on all data collected and test for differences between newly and initially estimated parameters. These methods include the Lagrange Multiplier (LM) test statistic (Glas, 1998, 1999) and the Cumulative Sum (CUSUM) statistic (Glas, 1999; Veerkamp, 1996). The LM test statistic has the advantage of known asymptotic chi-square distribution while the critical value for the CUSUM statistic is to be determined empirically in practical situations (Glas, 1999). However, both indexes require re-calibration of the item parameters, which could be a big challenge for most CAT programs. Unlike paper-and-pencil tests, a CAT item is not delivered to all test takers, but rather targeted to examinees within a narrow range of ability levels. As a result, in order to identify misfit items using the above-mentioned methods, it might take a long period of time to

accumulate a sufficiently large number of examinees with a wide range of ability, which is quite inconvenient, and sometimes impossible, in practical settings.

Bock, Muraki, and Pfeiffenberger (1988) proposed a statistical procedure for detecting item parameter drift in item pools for long-term testing programs. Their interest focused on the item content and secondary school curricula shifts.

Isham and Donoghue (1998) used Monte Carlo method to evaluate several indicators of item parameter drift in a simulation study. Three types of indicators were used in their study to detect drift: (1) IRT-based measures; (2) Mantel-Haenszel based measures; and (3) BILOG/PARSCALE Item-level χ^2 statistics. Overall, they concluded that Lord's chi square (Lord, 1980) measure was the most effective in identifying items that exhibited drift.

A second type of continuous monitoring method analyzes item performance based on discrete intervals. The z_c statistic described by Smith, Wang, Wingersky, and Zhao (2001), is calculated at sequential time intervals following each occasion on which an item is administered to significant numbers of examinees. This statistic is based on comparing observed and expected numbers correct within each time-based examinee sample. Again, observed values are based on some initial estimate of an item's performance.

Glas (2000) discussed a discrete approach to monitoring item parameter drift and differential item functioning. This method, called the Lagrange Multiplier test or efficient score test, is like that of Bock, Muraki and Pfeiffenberger (1988) in that it compares a restricted model that constrains item performance as stable over time with a general model that allows performance to drift.

Davey and Stone (2005) proposed a trend model to monitoring change that is based on modeling trends in item performance over time. The trend model assumes that over time (and perhaps repeated exposure) an item will change gradually. The trend model begins by assuming that examinee responses conform to the 3-parameter logistic model. Performance trends are then modeled by allowing item parameters to change as a function of time. Denote the initial (assumed) item parameters as a_0, b_0 and c_0, and let m represent time or the discrete occasions on which the item was administered. Then one simple trend model gives the item discrimination and difficulty parameters on each occasion m as the following:

$$a_m = a_0 e^{\alpha m}$$

$$b_m = b_0 - \beta m$$

where α and β are the trend parameters to be estimated. The "trended" item parameters a_m and b_m are those that are predicted to apply on each occasion. This model, simple as it is, is flexible enough to characterize widely different aspects and rates of change.

An obvious shortcoming of most statistics is that a re-calibration is needed in most cases. Practically, large CAT programs usually assemble item pools several months ahead of test administration. Some attractive items (e.g., items with high information) could appear in multiple pre-developed pools. Early detection of items with substantial item parameter drift, especially those compromised items, could help testing programs take appropriate early action, such as blocking the problematic items from active use or removing them from subsequent pools. So, monitoring item behavior in a timely fashion becomes an extremely important practical issue for the CAT programs. The re-calibration oriented methods are not able to meet this need in practice.

However, research on item parameter drift has found that naturally occurring amounts and magnitudes of drift tend to have a very minor impact on the resulting ability distribution. Wells, Subkoviak, and Serlin (2002) investigated the effect of item parameter drift on ability estimates under IRT. They simulated item response data for two testing occasions for the two-parameter logistic model under several crossed conditions. Their results showed that item parameter drift under the simulated conditions had a small effect on ability estimates. Even when *a* and *b* parameters were increased by .5 and .4, respectively, for 20% of the items, θ estimates were expected to deviate on the two tests by no more than 0.14 logits, for any true θ value. Similarly, Rupp and Zumbo (2003a, 2003b) found that examinees' scores were changed only slightly, unless the amount of simulated item parameter drift was unusually large.

2.5 Item Disclosure Detecting Model

McLeod, Lewis, and Thissen (1999) proposed a Bayesian method for detecting item pre-knowledge in CAT. When test takers use pre-knowledge of the items their item responses are likely to deviate from the underlying IRT model and estimated ability may be inflated. This deviation may be detected through the use of person-fit indices. They proposed a Bayesian log odds ratio index, which is much like the concept behind optimal appropriateness indices developed by Drasgow and Levine (1986). In the posterior log odds ratio approach to person-fit, c represents the dichotomous item pre-knowledge state (c and $\bar{c}$). If the state is c, then the test taker's response pattern is "nonfitting" and the test taker has not memorized any of the items and he or she is using his or her underlying proficiency to respond to the test. The probability p(c) that a test taker is using item pre-knowledge is updated after each item response. These "item pre-

knowledge" probabilities are based on the IRT parameters (assumed unknown), a model describing the probability of item pre-knowledge, p0(c) is a specified value that reflects the expected proportion of test takers believed to be using item pre-knowledge. This number may be established using empirical evidence from traditional approaches to detect cheaters, or prior elicitation based on the decision theory literature.

Segall (2001) presented a method for assessing consistency of test performance across two occasions, where on one occasion the level of performance may be misrepresented, and on the second occasion it is not. This method is based on the application of Bayesian model assessment methodology to multidimensional item response theory. His concern originally stemmed from the Internet based exam or some similar situation where the test takers were not sufficiently monitored. His solution was to administer a short second exam given under secure, proctored conditions. If the performance levels on the initial exam are consistent with the short proctored verification exam, then the first exam score is verified otherwise the initial test-scores are invalidated and the test taker is required to retake an alternate (the third) exam under proctored conditions. This procedure can be a complementary one in some practical high stake testing programs. The item pool in these programs can be divided into several sub-pools and different exposure rate can be set to different pools.

Segall (2002) developed an item response model for characterizing test-compromise that enables the estimation of item-preview and score-gain distributions observed in CBT. Markov-Chain Monte-Carlo (MCMC) was used to estimate the model parameters and posterior distributions. His simulation study showed that the model did provide useful summaries of test-compromise and its impact on test scores.

Segall (2004) took an additional step. A new sharing item response theory (SIRT) model was presented that explicitly models the effects of sharing item content between informants and test takers. This model is used to construct adaptive item selection and scoring rules that provide increased precision and reduced score gains in instances where sharing occurs. The adaptive item selection rules are expressed as functions of the item's exposure rate in addition to other commonly used properties (characterized by difficulty, discrimination, and guessing parameters). Based on the results of simulated item responses, the new item selection and scoring algorithms compare favorably to the Sympson-Hetter exposure control method. The new SIRT approach provides higher reliability and lower score gains in instances where sharing occurs.

This model expands the standard IRT model to incorporate the possibility that a correct response to an item can be influenced by one of three sources: the examinee ability level, guessing, and sharing item content among examinees. According to the model, an examinee can be characterized by two parameters, one continuous and one discrete. The continuous parameter θ denotes ability level. The discrete parameter h (where $h = 0, 1, 2, \ldots, n_h$) denotes the number of informants that have taken the test previously and shared some or all of these items with the examinee. Each item administered to an informant is assumed to be disclosed to the examinee with probability φ, and all items disclosed by the informant(s) are assumed to be answered correctly by the examinee if contained in their test. Item i is characterized by a known exposure parameter $e_i = \tau_i$ (where τ_i is defined as the probability of receiving item i) and by known discrimination, difficulty, and guessing parameters a_i, b_i, and c_i,

respectively. Then according to the SIRT model, the probability of a correct response to item *i* conditional on parameters θ and *h* is given by

$$p_i\left(u_i = 1 \mid \theta, h\right) = 1 - \frac{\left(1-c_i\right)\left(1-\varphi e_i\right)^h \exp\left(Da_i b_i\right)}{\exp(Da_i b_i) + \exp\left(Da_i \theta\right)}$$

where $u_i = 1$ and $u_i = 0$ denote correct and incorrect responses to item *i*, respectively. Note that 2-15 reduces to the standard three-parameter logistic model (Birnbaum, 1968)

$$p_i\left(u_i = 1 \mid \theta, h = 0\right) = c_i - \frac{\left(1-c_i\right)}{1 + \exp[Da_i\left(\theta - b_i\right)]}$$

for the case where the examinee benefits from zero informants (i.e., $h = 0$). Also note that when $\varphi e_i > 0$, that $p_i\left(u_i = 1 \mid \theta, h\right) \to 1$ as $h \to \infty$. That is, when φe_i is greater than zero, the conditional probability of a correct response approaches 1 as the number of informants approaches infinity.

Chang and Zhang (2002, 2003) developed an item pooling index to assess CAT item pool security. Their work demonstrated how important the problem of the security of CAT is theoretically. Before their study, an item overlap rate for a group of examinees had been defined as the ratio of the expected number of overlapping items encountered by two randomly sampled examinees from the group over the test length. The item overlap rates were computed by the percentages of the items that are shared by pair of examinees and then averaging across the pairs of examinees from the group. This item overlap index is referred as average item overlap rate in Way (1998). However, when there are more than two pairs of examinees, test overlap rate that is defined as the ratio of the expected number of overlapping items encountered between two randomly sampled examinees and the test length, has some limitations. It only

considers common items shared by two examinees, and it does not distinguish between beneficiaries and non-beneficiaries.

Chang and Zhang (2002) derived some indices to compute the degree of what they called item sharing and item pooling in a random sample of examinees. Item sharing implies that the items have been known by a group of examinees should be considered useless since too many examinees have prior knowledge. It represents the concept of the number of overlapping items shared among a group of randomly selected examinees. Item pooling, on the other hand, implies that one future examinee is able to gather information from several people who have taken the test. Therefore, item pooling is the number of overlapping items between one examinee and a group of examinees. Chang and Zhang (2002) pointed out that the threat to CAT was item pooling instead of item sharing. With this idea in mind, they derived the theoretical distribution of the information sharing variable and information pooling variable. Their conclusion is pessimistic. Assuming the test length is 30 and the item pool size is 700, if a thief, who is usually sent by coaching schools, is able to memorize 20 items, at most 20 thieves are needed to steal about 60% of the item pool.

Based on the theoretical derivation of Chang and Zhang (2002), Yi, Zhang, and Chang (2005) developed a computer program to examine the relationship among item pool size, the number of items each examinee can memorize, and the percentage of the item pool that can be compromised by sending a group of professional test takers, who take the test to memorize items and then share the information with others. The analytical results indicate that an operational CAT item pool needs to include, among other considerations, a large number of items. Test security can be strengthened by

increasing the size of an item pool from several hundred to a few thousand items. The results presented in their paper are based on the randomized item selection procedure that has the best test security control. However, even under the best test security control situation, only a handful of professional test takers are needed to compromise a sizable portion of an item pool when only several hundred items are in the item pool. In practice, the randomized item selection method is rarely used in actual CAT programs. The often-used maximum item information selection method is known to result in a skewed item usage in a pool. Therefore, the test security concern for a real CAT program is greater than that shown in the paper. The results of this paper indicate, in addition to look at the likelihood of including more items in a pool, the possibility of applying methods other than a maximum item information selection procedure in CAT needs to be explored.

2.6 Answer Coping Indices

Some methods have been developed to identify examinees who engaged in copying answers from others taking a test (Assessment Systems Corporation, 1993; Bellezza & Bellezza, 1989; Hanson, Harris, & Brennan, 1987; Holland, 1996; Sotaridona & Meijier, 2003; Wolleck, 1997). According to Cizek (1999), most of these methods can be classified into two types. One type of method compares an observed pattern of response to a known theoretical distribution. In the second type, the probability of an observed pattern is compared with a distribution of values derived from independent pairs of examinees who took the same test.

Wolleck (1997) provided a ω index based on IRT, which is an example of the first type of the methods. Suppose examinee *c* copied some answers from examinee *s* on

a multiple-choice test with V options in each item and let h_{cs} be the item where response c matches response s. Given the response of s on item i is k, let $p_{ik}(\theta_c)$ denote the probability that examinee c selects the same option k on item i. Wolleck (1997) showed that the probability is given by

$$p_{ik}(\theta_c) = \frac{\exp(\zeta_{ik} + \lambda_{ik}\theta_c)}{\sum_{v=1}^{V} \exp(\zeta_{iv} + \lambda_{iv}\theta_c)}$$

where ζ_{ik} and λ_{ik} are intercept and slope parameters. The expected value and standard deviation of h_{cs} are

$$E(h_{cs} \mid \theta_c, \mathbf{U}, \boldsymbol{\zeta}, \boldsymbol{\lambda}) = \sum p_{ik}(\theta_c)$$

$$\sigma_{h_{cs}} = \sqrt{\sum p_{ik}(\theta_c)(1 - p_{ik}(\theta_c))}$$

ω is defined as the residual between the observed and expected value of h_{cs}. That is

$$\omega = \frac{h_{cs} - E(h_{cs} \mid \theta_c, \mathbf{U}, \boldsymbol{\zeta}, \boldsymbol{\lambda})}{\sigma_{h_{cs}}}$$

The larger the value of ω, the stronger the evidence of examinee c copied examinee s.

Holland (1996)'s K index is an example of the second type of these methods. In this method, number incorrect group $r = 1, 2..., c', ..., R$ is defined so that examinee($j = 1, ..., J$) has the same number of wrong answers. c' indicates group membership of c. The number of examinees in number incorrect group r is denoted by J_r so that $J_{c'}$ is the number of examinees with the same number of wrong answers as examinee c. Let U_{inj} be response of examinee j in number incorrect group r to item i and W_s be the set of items, of size of w_s, answered incorrect by examinee s. For each

examinee *rj* an indictor A_{irj} is defined as 1 if $U_{irj} = U_{is}$ and 0 otherwise. The number of matching incorrect answers of *rj* and *s*, denoted by M_{rj} is defined as

$$M_{rj} = \sum_{i \in W_s} A_{irj}$$

Let $M_{c'c}$ be the number of matching wrong answers between *c* and *s*. The probability of random variable $M_{c'c}$ can be computed by

$$p\left(M_{c'c} \geq m_{c'c}\right) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} p^w \left(1-p\right)^{w_s - w}$$

If the value of *p* is estimated from the observed data by $\hat{p} = \frac{\bar{M}_{c'c}}{w_s}$, where $\bar{M}_{c'c}$ is the means of $M_{c'c}$ and w_s is the number of wrong answers of the source, Holland (1996) defined the *K* index as

$$p\left(M_{c'c} \geq m_{c'c}\right) = \sum_{w=m_{c'c}}^{w_s} \binom{w_s}{w} p^w \left(1-p\right)^{w_s - w}$$

This is an upper-tail probability. It can be compared to a chosen significance level, such as 0.01. When the probability is less than or equal to the value the examinee *c* is identified having a pattern of responses unusually similar with examine *s*.

2.7 Conclusion

The literature review shows that though test security is a major concern of CBT not very many studies have focused on it. However, quite a few research directions are suitable to address the problem. For example, many approaches have been proposed to minimize the usage of each item (i.e. item exposure control) which intend to reduce the possibility than some items are seen more frequently than the others, to investigate

whether or not the assumed IRT model describes each examinee's behavior, and to identify an item whose parameters drift after a certain period of time. Substantially less attention has been paid to on-site monitoring of item behavior. Minimizing the possibility that an item is administered and spotting the item once it is exposed are two facets of one question. Everybody knows it is easy to compromise some frequently used items if the item exposure rates are not controlled. On the other hand, few people have noticed that if we cannot spot an item that has been known by a group of examinees before they take the test, we will never know how long an item should be used or how often an item pool should be rotated. This dissertation research will address the latter issue.

But, no matter how fancy these methods are, they cannot eliminate the possibility that some items are still stolen by some "professional thieves" or just shared by examinees. Although item exposure control has been a major concern in the developments and implementations of CBTs, however, unlike in many other aspects in CBT, there is a lack of theoretical development in limiting item exposure rate (Zhang & Zhang, 2002). In addition, Way (1998) pointed out that there is no common understanding as yet about issues such as what represents acceptable item exposure rates and how long an item pool should be used. For example, the desired maximum exposure rates were set at 0.20 in the case studies of the computerized adaptive test (CAT) versions of the GRE General (Eignor, Stocking, Way, & Steffen, 1993). But, how do we know an exposure rate is high or low? Why is it 0.2, and not 0.3?

In this case, another interesting research direction is whether or not we can develop some methods to detect an exposed item once it has happened (Lu &

Hambleton, 2003; Han, 2003). This direction reveals the other "facet" of the security problem. Can we raise the alarm to the test administrators if an item or some items are exposed? This kind of method cannot prevent an item from being exposed but the value is to let test administrators know whether they can still use an item, or if the item should be removed or blocked from the item pool.

CHAPTER 3

METHODOLOGY

3.1 Moving Average

A moving average system is widely used in the stock market to predict the trend of stock prices. The definition and computations of moving averages are:

Given a sequence of values $\{x_1, x_2, \cdots, x_t, \cdots\}$ and a window size $K > 0$, the k-th moving average of the given sequence is defined as follows:

$$
\begin{aligned}
y_1 &= \tfrac{1}{k}\left(x_1 + \cdots + x_k\right) \\
y_2 &= \tfrac{1}{k}\left(x_2 + \cdots + x_{k+1}\right) \\
&\vdots \\
y_{n-k+1} &= \tfrac{1}{k}\left(x_{n-k+1} + x_{n-k+2} + \cdots + x_n\right)
\end{aligned}
$$

Moving average is a form of averages that has been adjusted to allow the long-term trends of a time series data to be clearer. One property of the moving average is: moving averages with a small window size respond to the trend underlying the date series faster than the ones with big window sizes. That is to say, the bigger the value of k is, the more stable the moving averages are. When k becomes big, the change of one element or a few elements will not affect the moving averages significantly. On the other hand, when k becomes small, minor changes of a few elements or of even one element will result in immediately noticeable changes in moving averages. In the stock market, a series of moving averages with different window sizes of k are usually computed for a stock (e.g. 5 days, 10 days, 30 days, a half year, a year, etc.) to forecast

the changes of the stock prices. If the trend of the stock price is going up, the short-term moving averages will increase before the long-term moving averages do so as well.

When a variable, like the number of unemployed, or the score of examinees taking a CBT, is observed at a sequence of time intervals, this sequence of data is called "time series data". The essential difference between time series data and the random samples of observations that are discussed in the context of most other statistical methods is that data points taken over time may have internal structure (such as a trend or periodic variation) that should be accounted for. The underlying trend usually is difficult to see because of the presence of periodic variation and random error. Moving average is a powerful tool that can be used to eliminate these components and random errors from time series data.

Unlike in conventional paper and pencil tests, where a single test is administrated to a whole population of examinees at the same time, "time" is not an interesting variable. But, in the environment of CBT, a string of examinees take a test successively. Sometimes "time" may affect the test results dramatically (though we hope not!). Therefore, the scores of examinees or observed examinee performance are time series data. There may be trend, periodic variation, and other variations underlying the data. For example, if an item pool is over exposed, the later examinees may take advantage of the knowledge of item pool to get higher scores. There will be an increasing trend underlying the score series. The concept of "moving averages" appears to have relevance too.

Let's begin from an assumption that will define the simplest situation of CBT. If the examinees taking a CBT exam in a relatively short time period are treated as a time

series sequence $\{x_1, x_2, \cdots, x_t, \cdots\}$ where x_t is the ability of the t-th examinee, this sequence is stationary both in mean and variance. In other words, there should be no significant increasing or decreasing trend on this sequence. At this case, the scores of the examinees would look like they are being randomly drawn from a fixed distribution with a common location and common scale.

Under this assumption, once we obtain a sequence of the moving averages of the item scores for a given window size of k, if this item is not compromised, the moving averages of the item scores on this item should be stable except for random variation associated with the window size (see Figure 3.5). In contrast, if this item is compromised, since some examinees who have acquired pre-knowledge of the items before the test administration are more likely to be scored higher than they should be, an increasing trend will occur in the moving average sequence. If we look at the plot of several moving averages with different window sizes we will find that the short-term moving averages increase before the long-term moving averages do so as well (see Figure 3.6). In reverse, whenever we find the short-term moving averages increase dramatically while long-term moving averages remain stable we might infer that the item pool is over exposed if there is no evidence that the abilities of examinees are increasing.

Mathematically, a moving average of the item scores equals to the item p value estimated on this set of examinees. Therefore the moving averages sequence of the item scores will be called moving p values in this and the following chapters. Obviously, the assumption is too strong for many operational testing programs. Examinee abilities are very likely to increase or decrease over time. High ability examinees may show up in

the beginning of the examination period while low ability examinees may show up later, for example. The moving p value sequence is dependent on the ability distribution of the examinees therefore will be useless in detecting item exposure on this case because item performance differences and ability differences may be confounded. We need to look for some item statistics that are free of the ability distribution.

A natural item index is the item *b* value under the IRT framework (Hambleton, Swaminathan, & Rogers, 1991). Obvious if we compute the moving averages of the item *b* values the sequence should be stable if the item is not exposed to examinees. But monitoring the moving averages of the item *b* values will involve too many calibrations --we would have to recalibrate the item parameters after every administration --which makes this solution impractical. The following four item statistics, which can capitalize on the advantages of IRT but do not require recalibration, seem much more promising:

3.1.1 Item Residual

The probability that an examinee with a given ability level answers an item with given characteristics correctly is assumed as an underlying model--usually a logistic curve--in IRT. An item residual is the difference between observed item performance for an examinee with a given ability and his/her probability of answering this item correctly.

$$r_{ij} = x_{ij} - p_{ij}$$

where *i* denotes the item and *j* denotes the examinee. x_{ij} is the observed score of the examinee on this item. p_{ij} is the probability that the examinee answers the item correctly under the hypothesized IRT model. Figure 3.3 displays two residuals for two

examinees with different abilities. The value of the residual is positive when an examinee answers the item correctly and negative when an examinee answers the item incorrectly.

If we compute the moving average of the item residuals for a certain item and a group of examinees, the moving average sequence should be stable as long as the sample size is big enough. Actually, it is expected that the moving average sequence should be around zero. However, if an item is exposed, the chances of giving correct answers to the disclosed items have been increased because some examinees have prior knowledge to the items, especially for those examinees with lower ability levels compared to the difficulty levels of the items. Consequently, more examinees answer the items correctly than expected; therefore, the moving average of item residuals will have higher positive values. This property makes item residuals potentially useful in detecting exposure.

The item residuals defined above are called raw item residuals (see, Hambleton, Swaminathan, & Rogers, 1991). A limitation of raw item residuals is that they do not take the sampling error into account (Hambleton, Swaminathan, & Rogers, 1991). The numerical value of the statistic depends on the trait level and therefore the raw residuals have different meanings along the trait scale. Therefore, some more commonly used statistics are raw residuals after some kind of standardization

3.1.2 Standardized Item Residual Z_c

Standardized item residual is the ratio of the raw residual over the standard error. Wang, Wingersky, Steffen and Zhu (1998) suggested using the following Z_c index to monitor the fitness of the item and model in a CAT operation.

$$Z_c = \frac{\sum_{j=1}^{N} (x_{ij} - P_{ij})}{\sqrt{\sum_{j=1}^{N} P_{ij}(1 - P_{ij})}}$$

where

x_{ij} is the binary score variable for item i of examinee j,

P_{ij} is the probability that examinee j, with the latent trait θ_j, gives a correct answer to item i,

N is the sample size on which the index is computed.

Suppose that there are a group of examinees, Z_c is computed to this group of examinees. Then by adding a new examinee to the sample and dropping the oldest one a new value is obtained while the sample size keep the same. The Z_c sequence we obtain should stabilize according to the definition of the statistic.

3.1.3 z_2 and z_3 (Zhu, Yu & Liu, 2002)

Smith, Wang, Wingersky and Zhao (2001) pointed out that index *Zc* is directed to a deviation between the observed overall number of right and the expected overall number of right. It measures an average deviation between the observed item response function and the expected response function across a pre-defined ability range.

Therefore, index *Zc* can be applied to uniform deviation only. To overcome the limitation of Z_c, Zhu, Yu & Liu (2002) proposed two new statistics z_2 and z_3.

In contrast to the overall difference between the observed and the expected total numbers of right among all examinees within a certain ability range, the computation of Z_2 first classifies examinees into K different ability groups ($k = 1, \ldots K$), then computes the weighted root-squared difference between the observed and the expected total numbers of right among examinees within the same ability group k, and, finally, sums the weighted differences across all K ability groups. The computation of Z_2 is given as,

$$z_2 = \frac{\sum_{k=1}^{K} (\frac{n_{ik}}{N_i}) \sqrt{(O_{ik} - E_{ik})^2}}{\sqrt{V_i}}$$

where

$$O_{ik} = \sum_{j=1}^{n_{ik}} u_{ijk}$$

u_{ijk} is the response (0,1) of examinee j in ability group k to item i, n_{ik} is the total number of examinees within ability group k, and N_i is the total number of examinees responding to item i; and

$$E_{ik} = \sum_{j=1}^{n_{jk}} P_i(\hat{\theta}_{jk})$$

is the 3P logistic function while $\hat{\theta}_{jk}$ is the CAT estimated ability for examinee j in group k; and Vi is the same as defined previously.

Index z_3 employs the conditional error variance within each ability group k, instead of using the grand error variance based on all examinees. In other words, z_3 first computes the weighted standardized difference within each ability group, then sums across all K ability groups. The computation of z_3 is shown as follow,

$$Z_3 = \sum_{k=1}^{K} \left(\frac{n_{ik}}{N_i}\right) \sqrt{\frac{(O_i - E_i)^2}{V_{ik}}}$$

and

$$V_{ik} = \sum_{j=1}^{n_{ik}} P_{ik}(\hat{\theta}_{ik})(1 - P_{ik}(\hat{\theta}_{ik}))$$

3.2 Statistical Control Chart

Statistical control chart is a statistical approach to monitor process variation for the purpose of improving the effectiveness based on continuous monitoring of process variation. Control charts are widely used to routinely monitor quality in engineering and manufacturing industry and are not relevant for paper and pencil tests. However, since in most cases examinees take a CBT in sequence, sequential monitoring some statistics of the test over time may prove to be necessary. For example, van Krimpen-Stoop and Meijer (2000) introduced using statistical process control techniques to detect person misfit in CAT.

One basic type of control chart is a univariate control chart, which refers to a graphical display of one quantity characteristic. If a single quantity characteristic has been measured or computed from a sample, the control chart shows the value of the quantity characteristic versus the sample number or versus time. In general, the chart

contains a center line that represents the mean value of the quality characteristic for the in-control process. Two other horizontal lines, the upper control limit and the lower control limit, are shown on the chart. These control limits are chosen so that almost all of the data points will fall within these limits as long as the process remains in-control. Figure 3.4 illustrates this point.

3.3 Determination of Control Limits

The control limits in the graph are usually set up in order that the probabilities of data points falling above the upper limit or below the lower limit would be very small. In the engineering world of the US, whether X is normally distributed or not, it is an acceptable practice to base the control limits upon a multiple of the standard deviation. Usually this multiple is three and thus the limits are called six-sigma limits. This term is used whether the standard deviation is the population parameter, or some estimate, or simply a "standard value" for control chart purposes.

In our case, if the null distribution of the item statistic is known, these limits can be set up theoretically by computing certain percentiles of the distribution. If the null distribution is unknown these limits can be set up by simulation or historical data empirically. For a sequence of examinees with know ability parameters the examinees' scores on each item can be simulated under a hypothesized IRT model and the item statistics introduced above can be computed and the obtained sampling distribution can be used to set up the control limits. It is a common practice in most of test agencies that an item pool is simulated on a theoretical assumed or historically obtained examinee population before it is packaged and deployed into test sites. Therefore the control limits can be set up in this phase. The control charts are then produced in live test

administrations. If the monitored sequence falls outside the control limits, we assume that the process is probably out of control (i.e. the item is exposed). Nevertheless, this may not mean that when all points fall within the limits, the process is in control. If the plot looks non-random, that is, if the monitored sequence exhibits some form of systematic behavior, there is still something wrong. To be sure, "in control" implies that almost all elements of the sequence are between the control limits and they form a random pattern.

In most cases, the determination of the control limits is a matter of professional judgment. On the one hand, using a too large control limits would run the risk of letting items with serious item exposure undetected. On the other hand, if the control limits are too small, many items with small or moderate amounts of deviation would be over flagged.

It is necessary to know the null distribution of the item statistic or make some assumptions to the distribution of the statistic one wants to obtain the control limits theoretically. Therefore, the determination of the control limits is impossible or inaccurate in some cases. An alternative method to monitor the trend of moving average is to plot several sequences of moving averages with different window sizes on one graph. This is commonly done in the stock market. Three is no sophisticated statistical model behind this method. The logic is short-term moving averages respond to the trend faster than does the long-term moving averages do so. In stock market, different people use moving averages in different ways. Here are two primary strategies that people use moving averages in the stuck market:

3.3.1 Filters

Filtering is used to increase your confidence about an indicator. There are no set rules or things to look out for when filtering, just whatever makes you confident enough. For example you might want to wait until a security crosses through its moving average and is at least 10% above the average to make sure that it is a true crossover. However, setting the percentile too high could result in "missing the boat". Another filter is to wait a day or two after the security crosses over, this can be used to make sure that the rise in the security isn't a fluke or un-sustained. Again, the downside is if you wait too long then you could end up missing some big profits.

3.3.2 Crossovers

Using Crossovers isn't quite as easy as filtering. There are several different types of crossovers, but all of them involve two or more moving averages. In a double crossover you are looking for a situation where the shortest moving averages crosses through the longer one. This is almost always considered to be a buying signal.

The test administrators can monitor the moving averages sequences in a similar way. Empirical filters and different levels of an alarm system can be set up after a large number of administrations. By this method, the test administers are aware of the trend of the score changes. They can determine the effect of any outside event to the test and take proper actions. Crossover is not very interesting in our application but a strong signal should be noticed when all short moving averages are crossing over the long moving averages.

3.4 Strategies for Dealing With Exposed Items

When an item is suspected to be exposed, a natural strategy is to remove or block it from the item pool. In practice this can be done by modifying the item exposure control model. If an item is flagged, the probability to administer this item can be set to zero so that this item will no longer appears in the future test. More ingeniously, the item exposure parameter can be linked to the monitored item statistics so that suitable actions can be taken before an item is flagged.

3.5 Simulation Studies

Several simulation studies have been conducted to investigate the usefulness of the proposed statistics. Before the design of the simulation studies, it is necessary to know which variables would be expected to affect the detection of the exposed items.

Several item delivery mechanisms have been widely. Linear fixed test (LFT) is considered less advanced to other item delivery mechanism while CAT is a very popular one. Recently, multi-stage test (MST) design is gaining more attention. However, the computation of moving averages is conducted to administrations for each item. Therefore, they are free of item delivery mechanisms. The simulation studies here were conducted on a linear fixed test, which consisted of 75 items. The parameters of the items are generated according to the following rules:

$$a_i \sim N(1, 0.2)$$

$$b_i \sim U(-3, 3)$$

$$c_i \sim \mathrm{U}(0, 0.25)$$

Common sense told us more examinees will benefit if a hard item is exposed while only a few examinees will benefit if a very easy item is exposed. This is because few examinees know the answer to the hard so if they can take advantage of prior knowledge of the item the probability to answer this item correctly will increase. Meanwhile, if a large part of examinees know the correct answer of an easy item, the probability for them to answer this item correctly is already very high. Therefore, the first variable that affects the item exposure detection is the value of the item parameters. The first of 12 items will be assigned pre-determined parameters that cover most common ranges for item parameters.

As mentioned before, the proposed item exposure detection statistics monitor examinees' responses to an item over time. It is independent of the delivery mechanism of the test. A simple linear test design was used without loss of generality of the findings.

Another important variable that affects the detection process is the ability distribution of the examinees who take the test. If there is no trend among the ability sequence for the examinees, the exposure detection will be expected to be easy. When there is significant trend existing among the ability sequence, the situation becomes complicated. First of all, we will expect the detection statistic will be independent of the ability distribution. For example, item difficulty defined in classical testing theory is an effective detecting statistic when there is no trend in the ability distribution over time but if the abilities of the examinees become higher over time, the item difficulty statistic will show a trend too, but will not be revealing of exposure necessarily. Therefore, classical item difficulty is not an effective statistic in this instance.

Three different ability distributions were considered in the simulation studies: (a) the examinees were randomly generated from a normal distribution with mean of zero and standard deviation of one. That is, $\theta \sim N(0, 1)$; (b) the examinees will drift from a lesser ability group to a higher ability group. For this case, each examinee is drawn from a normal distribution with standard deviation of one but the mean of the distribution is a linear function of sequence number. To be specific, the following distribution was used to generate the examinees: $\theta \sim N(-1+i/2500, 1)$; and (c) the examinees abrupt shift from a normal distribution with mean of -1 and standard deviation of 1 to another normal distribution with mean of 1 and standard deviation of 1. That is, for the first half of the examinees, $\theta \sim N(-1,1)$; for the second half of candidates $\theta \sim N(1,1)$. In simulating drift, we assumed that the poorer candidates, generally, would take the test early (average ability = -1.0) and then gradually the ability distribution would shift from a mean of -1.0 to a mean of +1.0 by the end of the testing window. With the abrupt shift in ability distribution condition, the first 50% of the candidates were sampled from a N(-1.0,1), for the second 50% of the candidates, they were sampled from a N(+1.0, 1) distribution.

The number of candidates used in the study was 5000. There were two different instances that an item was exposed. For the first instance, we always assumed it began from the middle of the examinee sequence. That is, the examinees from the 1st to 2499th did not have any pre-knowledge to any item. By this design, the situation of an item is secured and that of an item is exposed can be contrasted. In real life, this instance occurs when some test takers memorize the items they meet when they take the test and put these items on the Internet. This is called "organized theft" in some of the

measurement literature. This situation has proven to be a huge threat to CBT in some Asian countries. The other instance is that one item is known by examinees gradually. This occurs when some test takers describe the items to their friends and the news spreads.

To conduct the simulation study, the probability to answer an item correctly for a certain examinee when the item is secure is computed by an IRT model. The 3P logistic model was used in this study. However, the more challenging task is to simulate the situation when an item is exposed. To the best knowledge of the author, most current models assume the correct probability is one for an examinee when the item is exposed. This may not be correct in real life. Usually, even though the examinees have some prior knowledge to some items they will never know if they will be administered these certain items for sure. Otherwise either the item bank size is too small or the item exposure control does not work at all. In the simulation studies conducted in this study, an item exposure simulation model was developed for simulating exposed items. The probability that an examinee answers an item correctly is computed by

$$P' = P + \rho(1 - P)$$

where: P: probability computed from the three-parameter logistic IRT model based on an examinee's ability level and item statistics.

ρ: a positive number $0 \leq \rho \leq 1$, was varied in the simulations, to reflect the item exposure model in place.

The equation is called the 'item exposure simulation model' in this study. When ρ is zero, the probability that an examinee answers an item correctly is the probability computed from the IRT model. This provides the baseline situation where the item is

secure. If ρ is one, the probability that an examinee answers an item correctly is one, which simulates an extreme situation where the item is exposed. Every examinee knows the correct answer to the item and scores one on the item. The magnitude of ρ reflects the extent to which the item is exposed.

This equation can be applied to the whole population of examinees or a certain percentage of the examinees. When it is applied to the whole population all examinees have some pre-knowledge about the item. When it is applied to a fraction of examinees only this part of the examinees has pre-knowledge about the item. In the simulation studies conducted during this study, different percentages were investigated to determine the power of different detection statistics. Usually, it is not realistic that all examinees know the answer to an item. For most exposed items only a certain percent of the examinees would have knowledge of the items.

This equation can also be applied to all items or some particular items. When it is applied to one or a small number of items, the ability estimates of the examinees can be expected to be more accurate than when it is applied to many items. It was expected that it is easier to flag the exposed item(s) when the number of the exposed items is not big, since an estimated ability parameter for each examinee needs to be used in most of practical detection statistics. This is similar to the detection of items that exhibit DIF. When a lot of items exhibit DIF the total score itself is biased so that it is questionable to use the total score as a criterion.

Overall, the interesting simulation variables are:

a) Ability distribution:

i) Normal;

ii) Drifting;

iii) Abrupt shift.

b) Extent to which an item is exposed:

$0 <= \rho <= 1$

$\rho = 0$ is a base-line situation where the item is secure.

$\rho = 1$ is an extreme situation in which every candidate answers the item correctly.

$0 < \rho < 1$ is a situation where examinee performance, relative to ability and item statistics, is increased to reflect the fact that some general information is being disseminated about the item which gives examinees a boost in their likelihood of success, but not a guarantee they will answer the item correctly.

c) Percentages of examinees who have pre-knowledge of the items.

d) Number of items that are exposed.

e) Choice of item exposure detection statistics:

i) Classical item difficulty (P value);

ii) Raw item residual;

iii) Standardized item residual Z_c;

iv) Standardized item residual Z_2 ;

v) Standardized item residual z_3;

f) The statistical characteristics of the items:

b = -1.0, 0.0, 1.0, 2.0

a = 0.40, 0.70, 1.20

These statistics were crossed to produce 12 item types to focus on in the research. These items were embedded into the 75-item test and appeared in positions 1 to 12 (without any loss of generality to the findings).

All together the combination of these variables would yield an astronomical figure so that it would make the research a mission impossible. Therefore the simulation study was divided into several steps with focus on different problems. What was learned from each step influenced the design and implementation of each later step:

The first simulation study was conducted using the classical item difficulty as the detection statistic. It is easy to know that classical item difficulty is not an effective statistic when there are ability changes within the examinee group over time. Figure 3.7 displays the moving averages of classical item difficulty of an item. The first graph of Figure 3.7 shows the situation when the examinee ability is distributed normally. The moving average sequence is flat over time. The second graph of Figure 3.7 shows the situation when the examinee abilities gradually increase over time. There is an obvious increasing trend as well along the moving average sequence. The third graph of Figure 3.7 shows another situation that the examinee abilities change over time where the ability distribution changed abruptly at a certain time point. At this case the moving average sequence shows the similar abrupt increase at a time point slightly later.

However, stable ability distributions do occur in a lot of educational and psychological measurements and classical item difficulty still is a widely used item statistic in practice. The advantage of the classical item difficulty is that it is model free, so no strict assumptions need to be made (as is the case with IRT). The results based on classical item difficulty are also easy to communicate between professionals and people without much training of testing theories. Therefore, it is useful to begin the study using classical item difficulty. The primary purpose of this first study was to uncover the stability of the moving average sequences on different window sizes. It was conducted with windows size of 50, 100, 200, 300, and 500. Only a normal ability distribution was used in this first study.

In addition to classical item difficulty, two IRT-based item statistics raw item residual and standardized item residual Z_c, were introduced into the second simulation study as the detecting statistic to demonstrate the properties of the moving averages of these item statistics in different combinations of the variables. First, the simulation was conducted to demonstrate the invariant properties of IRT-based item statistics for different ability distributions. At the first step in this second study, the value of ρ was set to zero, which implied that all the items were secure, and the windows size was set to 200. The result showed, not unexpectedly, that classical item difficulty is not an effective detection statistic when there was an instable ability distribution but both raw residuals and standardized residuals showed significant invariance for different ability distributions. It also showed that standardized residuals were more stable at different item difficulty and ability levels. In this study, ρ was also set to be 0, first to show how these item statistics performed when the items were secure. Then, ρ was set to be 1,

which refers to the situation that an item is totally compromised. With this case every examinee knew the correct answer to the item so the probability to answer it correctly was one. This is an extreme situation, one which we would never expect in real life so it is unnecessary to simulate it for all combinations of variables. But the result of this step was interesting which highlights the difference between raw item residuals and standardized item residuals. After this phase of the study, the most extreme values for some variables were dropped.

Study three was conducted using Z_c, z_2 and z_3 as detection statistics. These three statistics are all a kind of standardized item residual with some variations. It was conducted only to select meaningful combinations of the variables since the generalization of the founding may not be restricted by the variables. For example, if we find z_2 performances better than the other two statistics on window size of 300, the result should be true as well when the window size becomes 400. ρ was set to a none-zero positive value. This third study was used to investigate the capabilities of the three item statistics in detecting exposed items in different situations so empirical type I error rates and power of each statistic were explored. Some general conclusions and guidance for practical work were provided in this phase.

3.6 Indices and Plots

At this point, some results that were computed and the plots that are given in later chapters are defined and explained:

3.6.1 Empirical distributions of the moving averages when $\rho = 0$.

The simulation was replicated a great number of times when $\rho = 0$. For each replication we obtained a series of moving averages and these moving average sequences were used to set up the boundaries. For each item statistic under a given window size, the 2.5 percentile and 97.5 percentile were obtained for each replication. The mean of the 2.5 percentiles and the mean of the 97.5 percentiles were then obtained for the complete set of replications and they were used as the boundaries to flag the exposed items when ρ was not zero.

The empirical distributions when ρ was zero and non-zero were compared to demonstrate how differently they are when there are exposed items.

3.6.2 Detecting exposed test items

Under the no exposure condition, a great number of replications of the simulations helped determine the empirical sampling distribution of each of the item statistics after each item administration for each of the 12 item types. To be specific, 100 replications were carried out and the approximate 2.5, 97.5 percentiles were determined along with the mean of the 100 item statistics. What was used to approximate the percentiles were the mean plus two standard deviations and the mean minus two standard deviations. Figure 3.7 shows these values over many item administrations. These extremes were used in the flagging (i.e. detecting of exposed items). Whenever an item statistic exceeded these boundaries, either a type I error is made (if no exposure had been modeled) or exposure is detected (if exposure had been modeled).

A more formal explanation of what is happening is as the following. Given a sequence of examinees:

$$\{\theta_1, \theta_2, \ldots, \theta_t, \ldots, \theta_{5000}\}$$

where θ_t is the true ability of the examinee *t*.

For item *i*, the binary score for examinee *t* are obtained:

$$\{x_{i1}, x_{i2}, \ldots, x_{i5000}\}$$

The item statistics are then computed based on a sub-sequence of the examinees. For example: when windows size *k* equals to 100, the sequence of moving *p* values is:

$$\{p_{100}, p_{101}, \ldots, p_{5000}\}$$

where

$$\begin{aligned} p_{100} &= \tfrac{1}{100}(x_{i1} + \cdots + x_{i100}) \\ p_{101} &= \tfrac{1}{100}(x_{i2} + \cdots + x_{i101}) \\ &\vdots \\ p_{n-k+1} &= \tfrac{1}{k}(x_{i,n-k+1} + x_{i,n-k+2} + \cdots + x_{i,n}) \end{aligned}$$

The sequence of moving item residuals is:

$$\{r_{100}, r_{101}, \ldots, r_{5000}\}$$

where

$$\begin{aligned} r_{100} &= \frac{1}{100}([x_{i1} - prob\,(a_i, b_i, c_i, \theta_j)] + \cdots + [x_{i100} - prob\,(a_i, b_i, c_i, \theta_j)]) \\ r_{101} &= \frac{1}{100}([x_{i2} - prob\,(a_i, b_i, c_i, \theta_j)] + \cdots + [x_{i101} - prob\,(a_i, b_i, c_i, \theta_j)]) \\ &\cdots \end{aligned}$$

The sequence of standardized item residuals is:

$$\{sr_{100}, sr_{101}, \ldots, sr_{5000}\}$$

where

$$sr_{100} = \frac{\sum_{j=1}^{100}(x_{ij} - prob(a_i, b_i, c_i, \theta_j))}{\sqrt{\sum_{j=1}^{100} prob(a_i, b_i, c_i, \theta_j)(1 - prob(a_i, b_i, c_i, \theta_j))}}$$

$$sr_{101} = \frac{\sum_{j=2}^{101}(x_{ij} - prob(a_i, b_i, c_i, \theta_j))}{\sqrt{\sum_{j=2}^{101} prob(a_i, b_i, c_i, \theta_j)(1 - prob(a_i, b_i, c_i, \theta_j))}}$$

...

The following two item statistics can be obtained in similar manners. Every time an oldest element of the sub-sequence is dropped and a new element is added until the last element is added into the computation.

For each simulation, we can obtain one sequence for each item statistic. The simulation process was replicated 100 times. Therefore, for each item statistic we can obtain 100 sequences. Three new sequences for each item statistic are obtained and plotted: Mean, Mean + 2*SD, Mean – 2*SD. For example, for moving p values, the means of the simulations are:

$$\{\frac{\sum_{h=1}^{100} p_{h100}}{100}, \frac{\sum_{h=1}^{100} p_{h101}}{100}, \ldots, \frac{\sum_{h=1}^{100} p_{h5000}}{100}\}$$

where *h* stands for the *h*th replication.

This sequence is plotted in the middle of the plot and the dotted lines are (Mean + 2*SD) and (Mean - 2*SD). The vertical axis is the values of the sequence and the horizontal axis is the order of the sequence.

Figure 3.6 is an example of the detection plot for moving p value, residual, and standardized residual in which the item is exposed after the item has been administered 2500 times.

3.6.3 Item exposure detecting plots

The empirical method described above is sometime impractical. An alternative is set up straight line control limits arbitrarily. At this case, the determination of control limits is a matter of judgment, which is a decision problem instead of a statistical problem. In practice, several control limits can be chosen to set up different alarm levels like people in the homeland security department do. The test administers can take different actions to different alarm levels.

In our simulation study, the straight control limits were set up by taking an additional step to the empirical results obtained above. The mean of (Mean + 2*SD) and the mean of (Mean - 2*SD) across all examinee orders were computed and these means acted as the two control limits.

3.6.4 Average number of times of item administration after exposure.

Given the control limits obtained by theoretical computation or simulation, once the moving average sequence exceeded the upper limit the item was flagged. The number of times of item administration from the point where the exposure was introduced for the item to the point when it was flagged was computed too. The average

of these numbers are reported in tables like Table 3.1. This index indicates the sensitivity of the method.

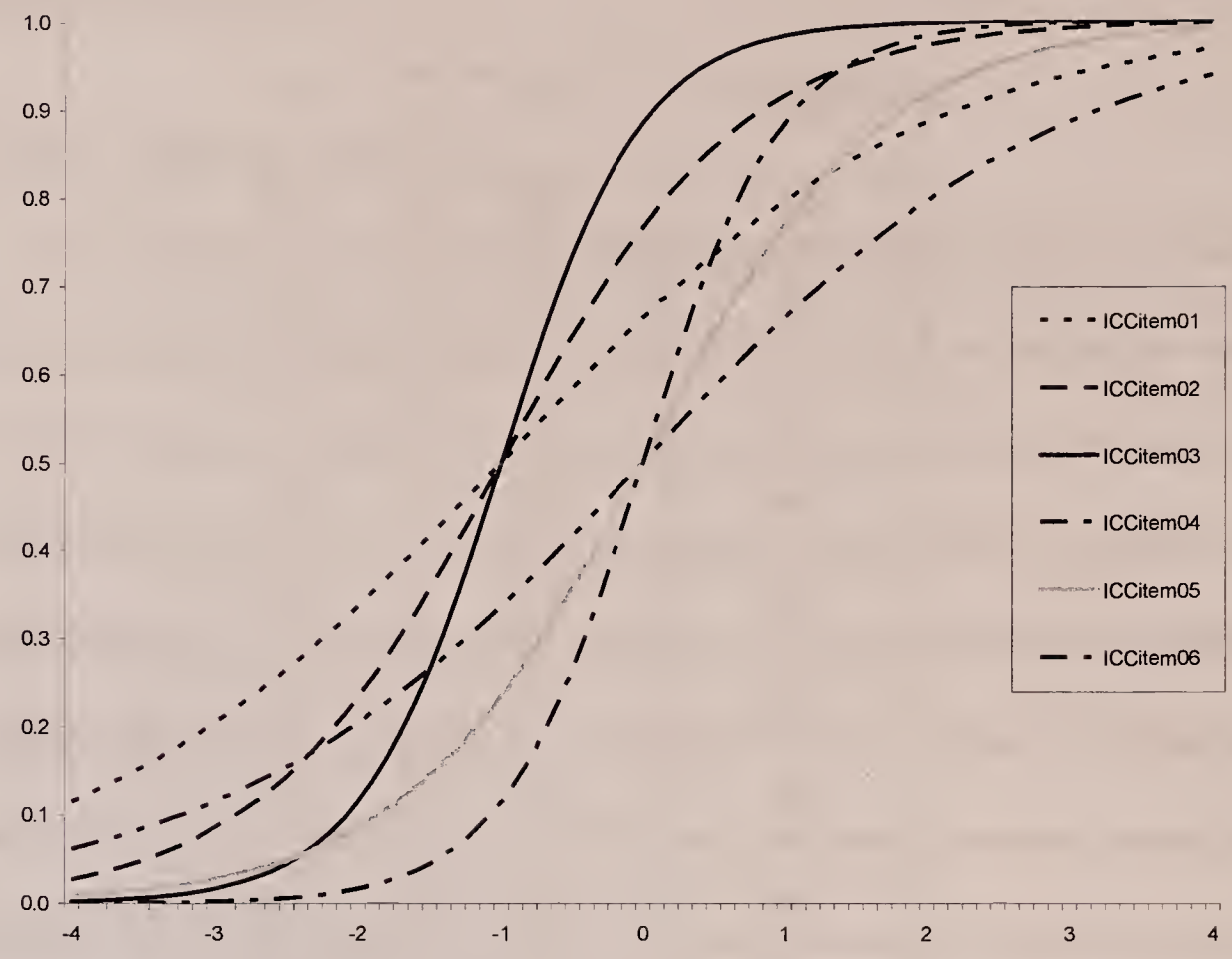
3.6.5 Type I error rate and power

The robustness of a statistical method is estimated by Type I error and power. Before applying the proposed method to practical situations we needed to estimate its type I error and power. In the simulation studies, when an item was exposed it always begins from the middle of the examinee sequences. That is: the 1st to 2500th examinees did not have any item pre-knowledge while the item exposure model took effect beginning with the 2501st examinee. To compute the type I error, we picked up the sub-sequence of examinees from 500st to 1549th and computed how many times the moving averages exceeded the upper limit. Since the item was secure, when the moving average exceeds the upper limit simply meant that a type I error was made. Similarly, the sub-sequence of the examinees from 3500st to 4499th was used to compute the number of times the moving averages was lower than the upper limit. Since the item was exposed every time the moving average was lower than the upper limit a type II error was being made. The power of the method is (1 minus type II error). For each item and combination of situations, the type I error and the power were tabulated.

Table 3.1. Item statistics of the 12 items to be studied

Item	a	b
01	0.4	-1.0
02	0.7	-1.0
03	1.2	-1.0
04	0.4	0.0
05	0.7	0.0
06	1.2	0.0
07	0.4	1.0
08	0.7	1.0
09	1.2	1.0
10	0.4	2.0
11	0.7	2.0
12	1.2	2.0

Figure 3.1. Item characteristic curves of the 12 items to be studied



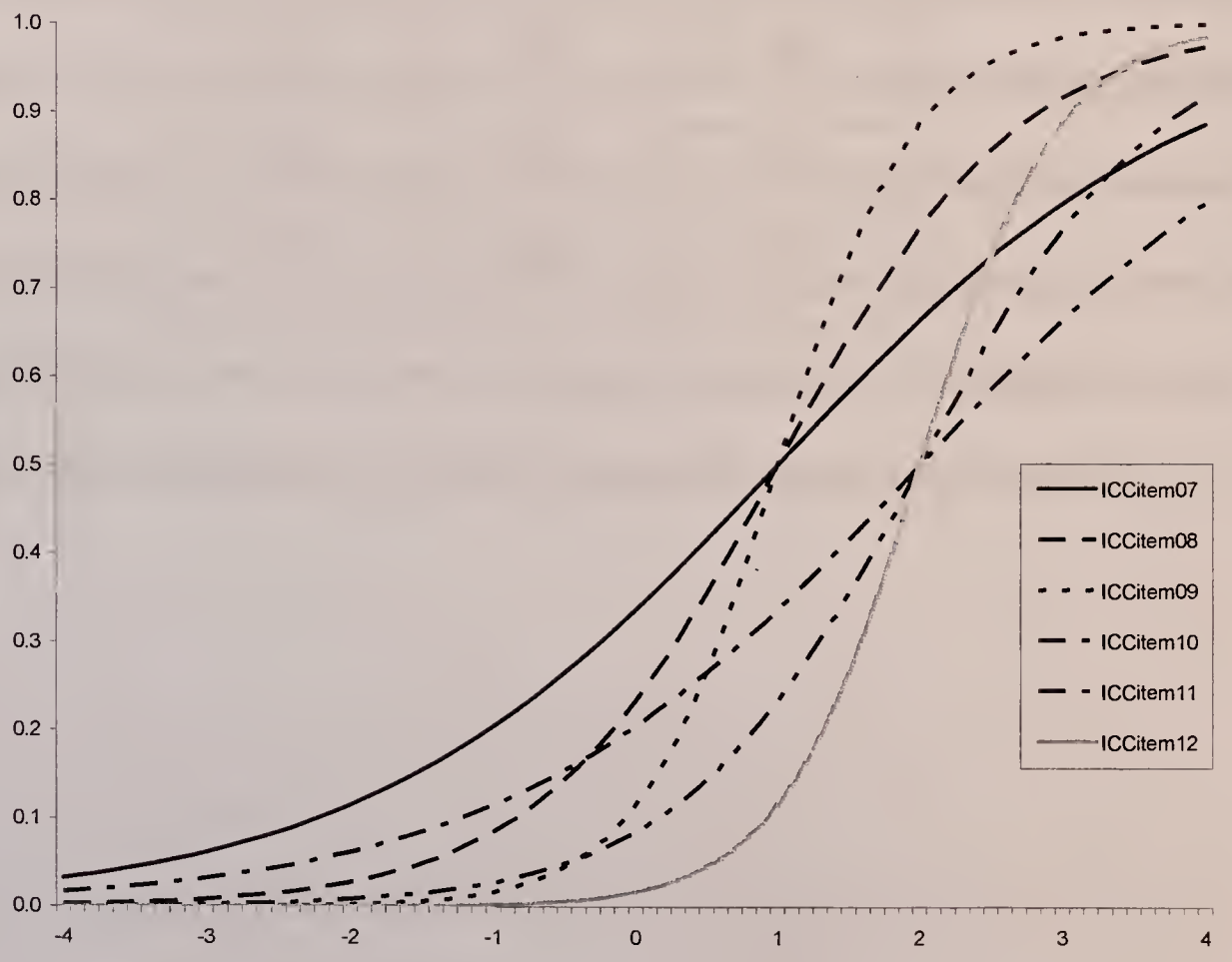
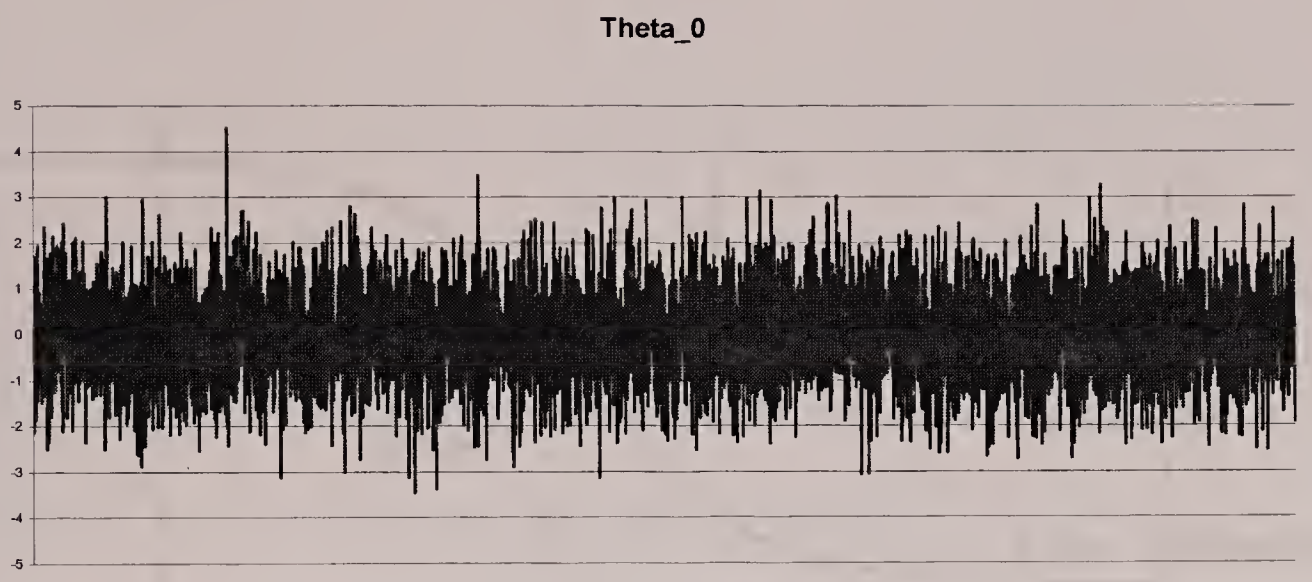
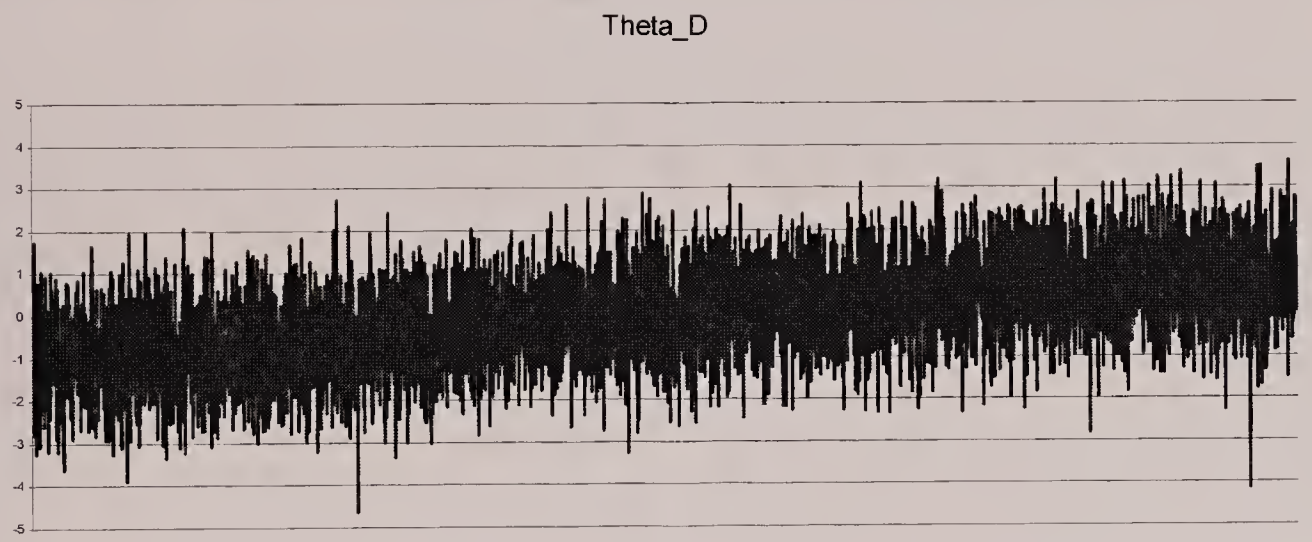


Figure 3.2. Stable, drift, and shift ability distributions





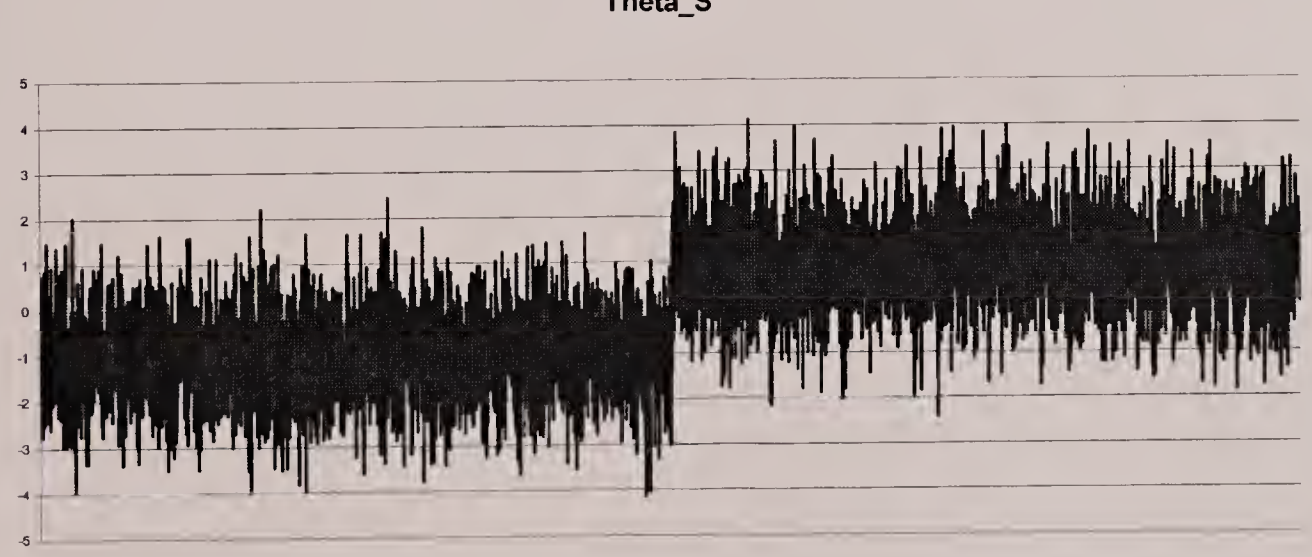


Figure 3.3. Display of item residual

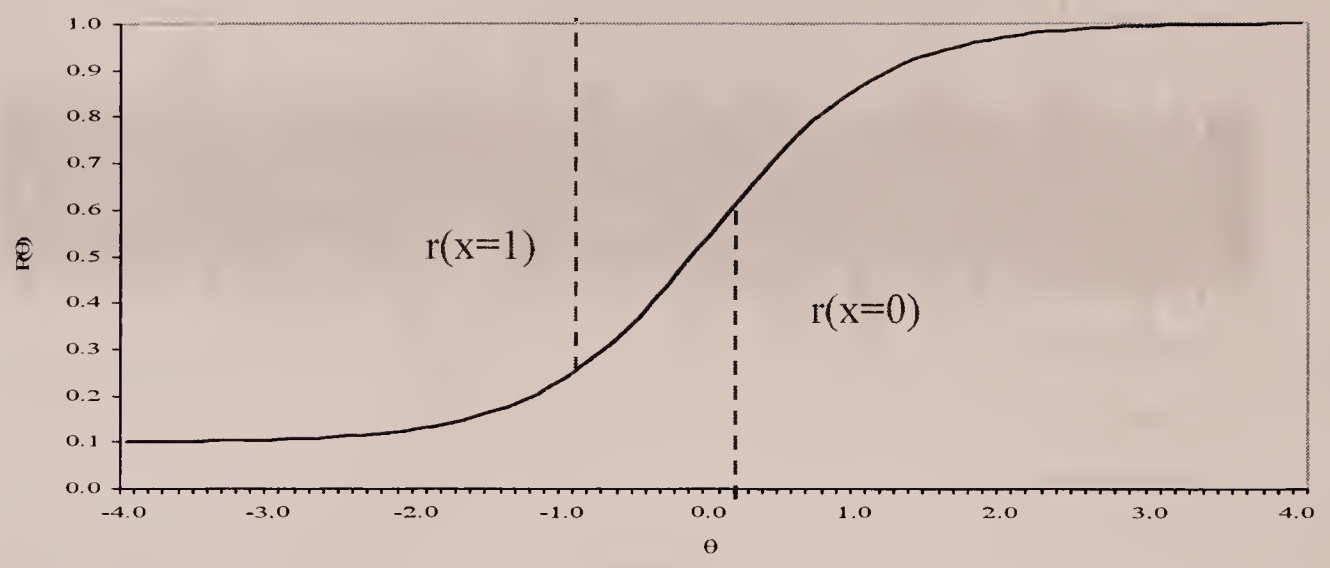


Figure 3.4. Statistical control chart

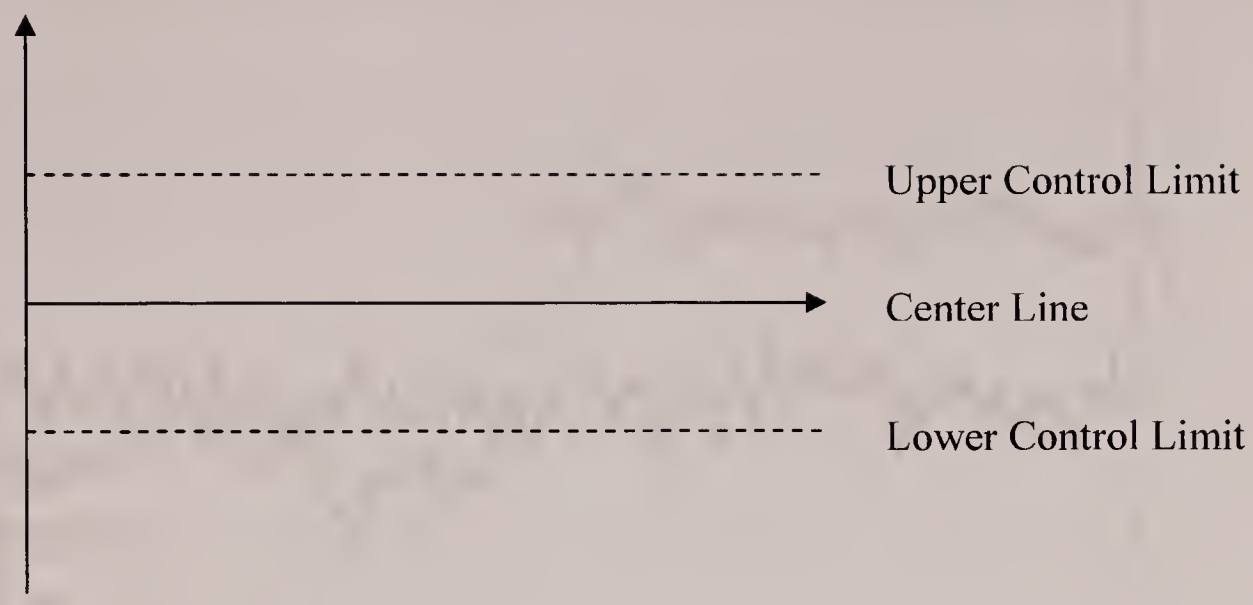


Figure 3.5. Moving averages with different window sizes when an item is secure

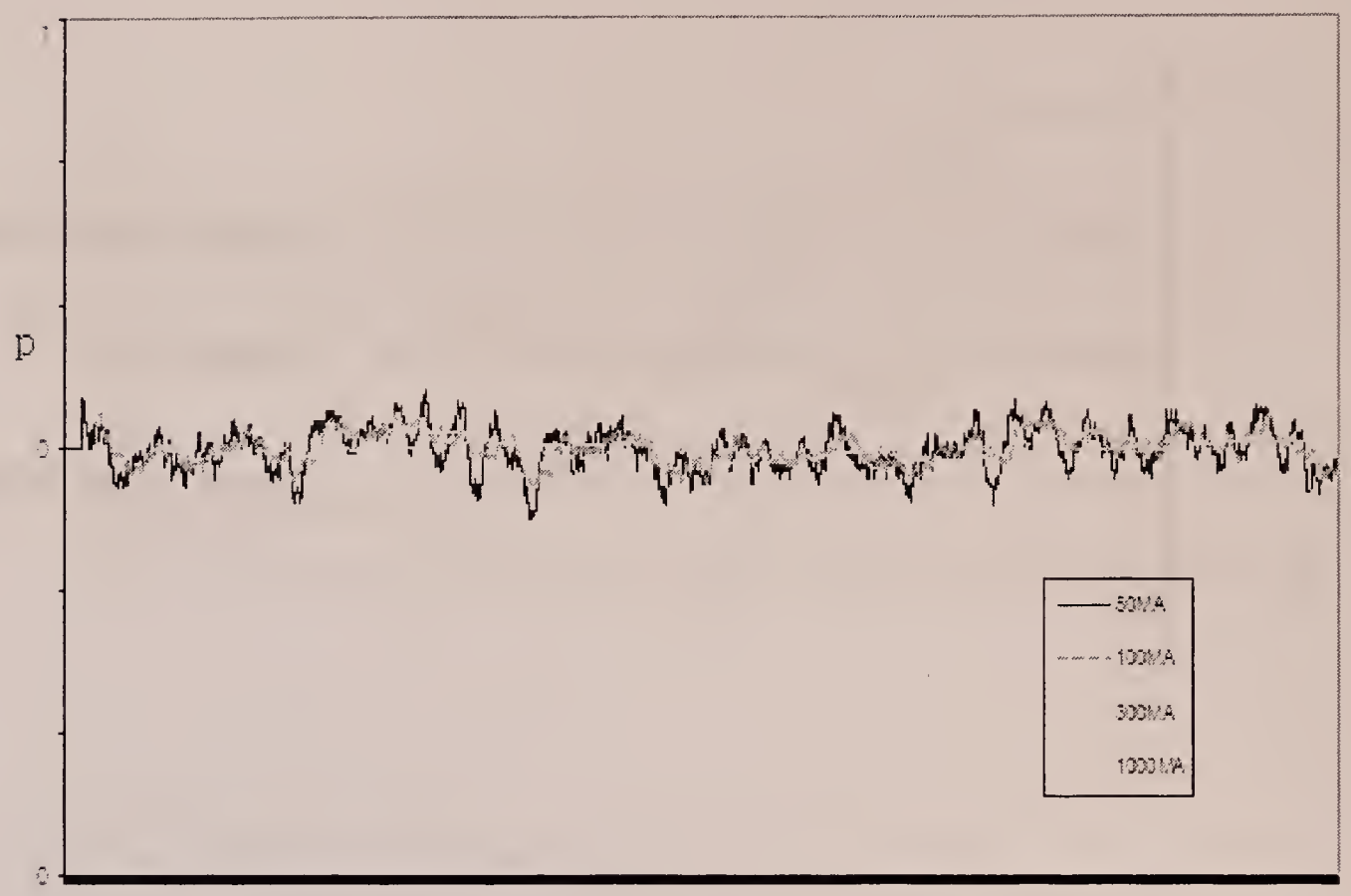


Figure 3.6. Moving p values with different window sizes when an item is exposed

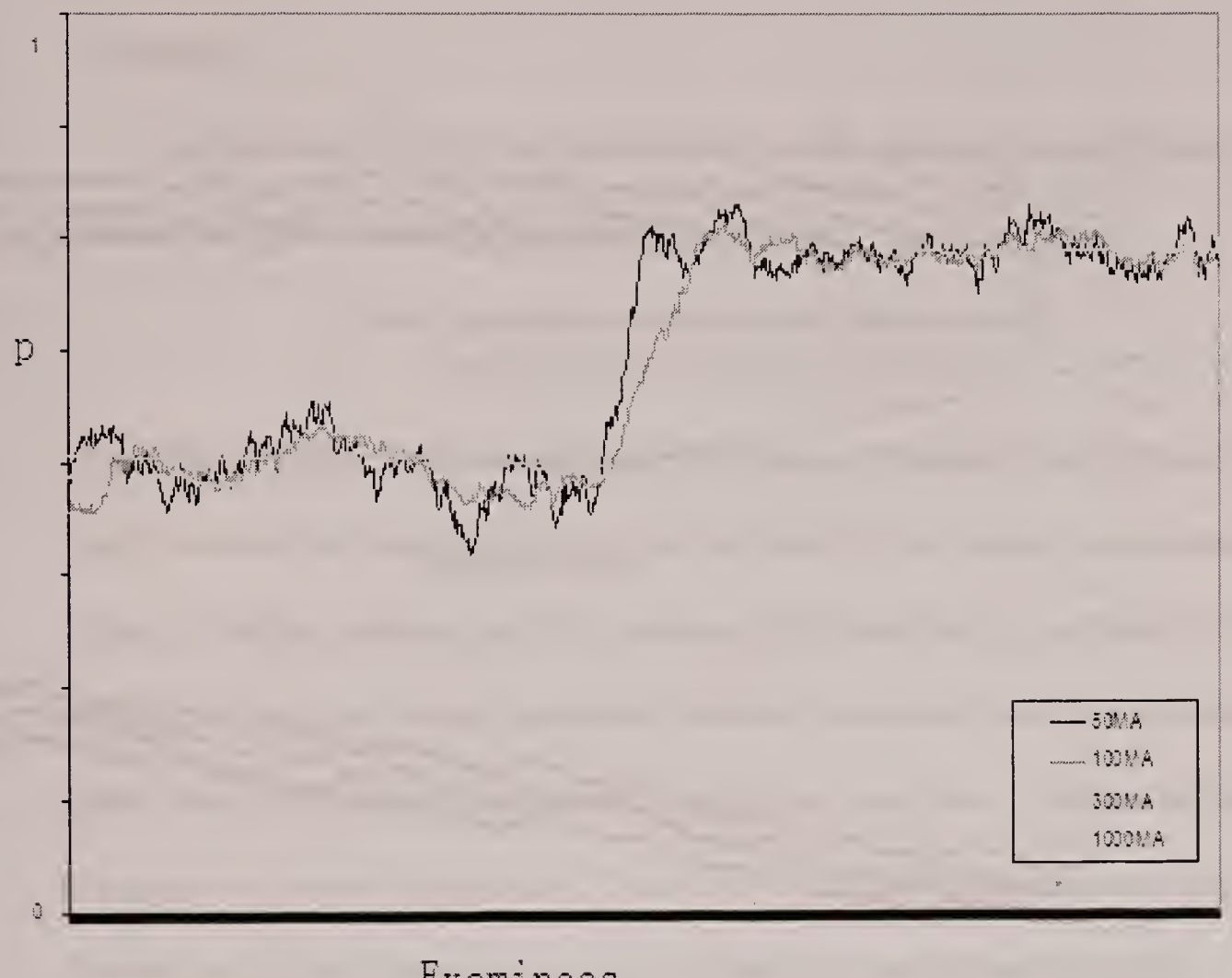
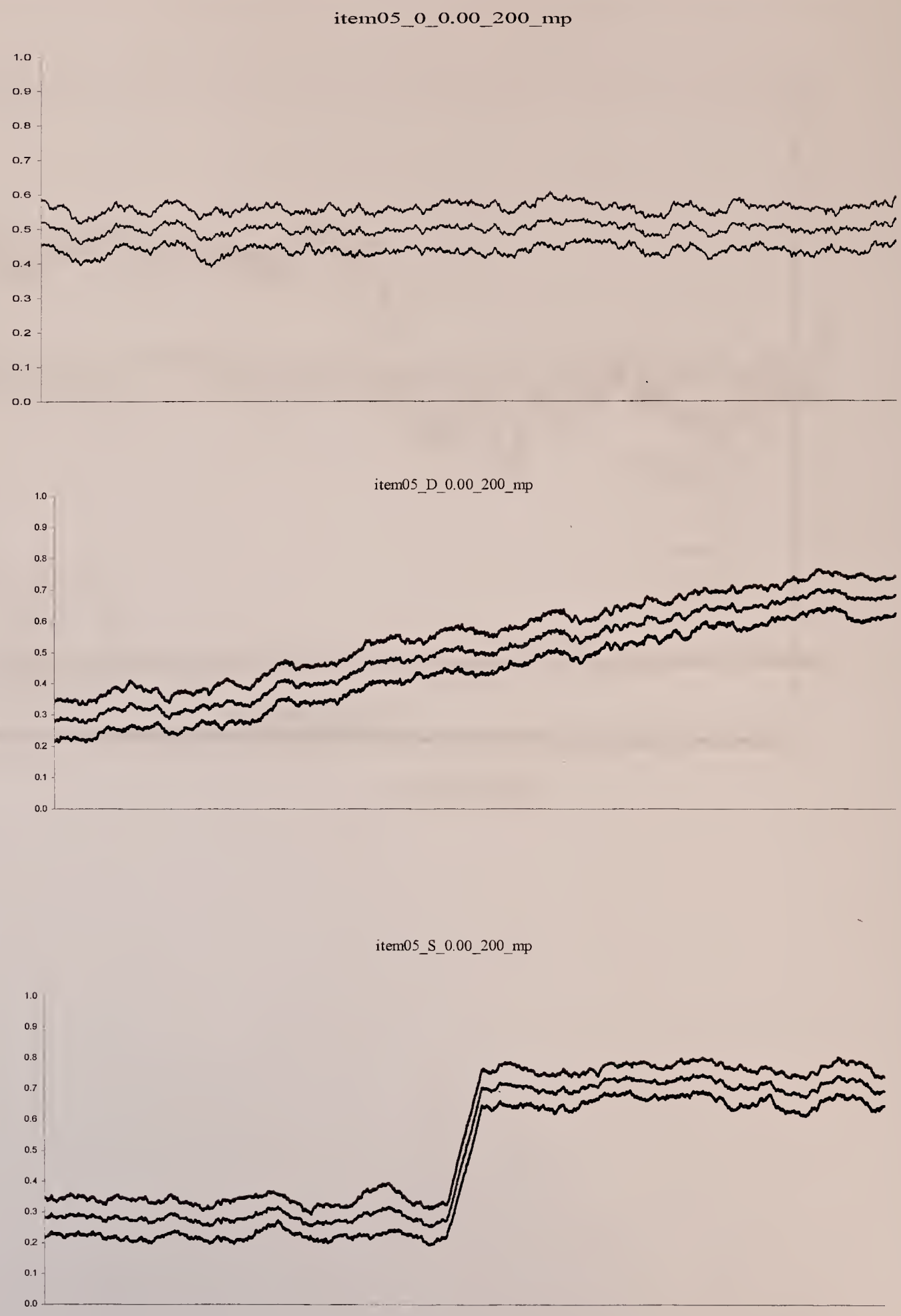


Figure 3.7. Moving p values for three different ability distributions



CHAPTER 4

SIMULATION STUDY 1

4.1 Purposes

In simulation study 1, the item statistic used for detecting exposed items is classical item difficulty, which is defined as follows:

$$p = \frac{\text{total examinees who answer the item correctly}}{\text{total examinees who attempt the item}}$$

As shown in Chapter 3, the moving average sequence of classical item difficulty was deeply affected by the distribution of examinee ability. This statistic is not effective if there is obvious evidence that the examinee ability distribution is affected by time. However, it is easy to use and understand when an assumption that the examinee ability is stable over time is reasonable based on the experience of the testing agency. It is especially useful operationally. It can uncover the information on an ability distribution and this information may be desirable for management purposes.

The simulation study described in this chapter was divided into two parts. The first part was used to develop an intuitive impression of how the moving average sequences perform when the window sizes vary. The finding of this part determined the window size used in the second part of the study. The second part was to show how this item statistic worked if the underlying ability distribution over time was stable.

4.2 Details of the Methodology

A fixed ability distribution was used in this study. Five thousand examinees were generated from a normal distribution with a mean of zero and a standard deviation of one. The whole set of 75 items was administered to these 5000

examinees and the probabilities for each examinee to answer each item correctly were obtained. For each item to be plotted, the moving p value sequence was obtained. Different window sizes from 50, 100, 200, 300, and 500 were applied.

First, ρ was set to be 0 which refers to the situation that the items are secure. This served two purposes: The simulation study was replicated 100 times so at each time point there are 100 values. The mean of these 100 values plus two times the standard deviation and the mean of these 100 values minus two times the standard deviation were displayed as well. The bands gives us a general idea how these moving average values fluctuate and, as discussed before, can be used to set up the control limits.

At the second part of the simulation study, ρ was set to be 0.25 and 0.50, respectively. The percentages of examinees with prior knowledge were set to one of two: 20 percent and 100 percent. This yielded four combinations: $\rho = 0.25$ and 20 percent, $\rho = 0.25$ and 100 percent, $\rho = 0.50$ and 20 percent, $\rho = 0.50$ and 100 percent. When an item exposure simulation model was employed it always took effect from the 2501st examinee. This part was done only on one fixed window size, which was100 in this study. This window size is relative small in practice so this study could be considered as a stress test. It would reasonable to be optimistic if it yielded good results.

For each combination, the whole test of 75 items was administered. The simulation study was not replicated in this case since both the number of examinees and the number of experiments we designed was very large. The control limits were set up according to the result of the first part of the study. They are the mean of the upper bands and the mean of the lower bands.

When an item was exposed, it always began from the 2501st examinee. In so doing, we could compare and contrast exposure and non-exposure in one simulation process.

4.3 Results and Findings

The first part of the simulation study was designed to look at the stability of the moving average sequences. Figures 4.1 to 4.10 plot the moving average sequences for all the 12 items that were monitored for different window sizes from 50 to 500 and no items were exposed. There are two plots for each window sizes. Figures 4.1 and 4.2 are for window size of 50, and so on.

Figures 4.1 to 4.10 show that for each item, the bands of the moving averages become smaller when the window sizes become bigger. This is reasonable since the moving averages become more stable when the window sizes become larger. For a given window size, since the distribution of the item responses is given by the Bernoulli distribution, the middle difficult items show the wider bands and the easier and harder items show the narrower bands.

Figures 4.1 and 4.2 show that window size of 50 was too small to obtain a stable moving p values. This was not a surprising result. Some fluctuations can still be observed when window size is 100 but they are noticeably flatter than windows size of 50 (Figures 4.3 and 4.4). The plots are very stable for window sizes over 200 (Figures 4.5 to 4.10). Therefore, from this perspective, 100 would appear to be a minimum acceptable window size for many studies to detect item exposure.

Figures 4.18 to 4.21 provide a general idea for how serious the extents of item leaking were under the situation of the four combinations of simulated variables. These four scatter-plots show examinee scores for two tests with and

without item exposure. Axis X is the scores on the test with no items exposed and axis Y is the scores on the test when item exposure exists.

Figure 4.21 is for the situation that ρ is 0.50 and 100 percent of the examinees have prior knowledge. It is clear from the plot that the examinees were obviously over estimated. This amount of over-estimation would be unacceptable in practice. Figures 4.17 and 4.18 are the detecting plots for the situation with $\rho = 0.50$ and 100 percent of the examinees having prior knowledge. These plots show that almost all items were clearly flagged while very few type I errors were made. Item 01 is the easiest item with poor discrimination. Even for this item there is no difficulty in flagging according to the second part of the plot while the first half of the plot is lower than the upper limit at most of the points, which indicated that there was very little possibility of committing a type I error.

Figure 4.18 shows that the trend to overestimate examinees was not very evident when ρ is 0.25 and 20 percent of the examinees had prior knowledge. This shows that the item leaking is not very serious in this case. Not surprisingly, it is the most difficult situation in which to flag the leaked items. In this situation, only item 11 and 12 (Figure 4.12) had high possibilities to be flagged while the type I error rate was low. These are extremely hard items with excellent discriminations. Item 10 was flagged with a weak signal but the opportunity to flag item 9 was greater than for item 10. There were some chances for item 6, 7 and 8 (Figure 4.11) to be flagged but it was also likely to commit a type I error. There were almost no chance to flag the other items.

Another two situations are between the two extremely situations. Some extent of overestimation in the ability scores can be seen from Figure 4.19 and 4.20.

When ρ is 0.50 and 20 percent of the examinees have prior knowledge, it was hard to flag item 1, 2, and 3 (Figure 4.13) which had the lowest value of the b parameter. But all other items were clearly flagged. When ρ is 0.25 and 100 percent of the examinees have prior knowledge, most of the items were flagged (Figure 4.15 and 4.16).

To sum up, when 100 percent of the examinees have prior knowledge to exposed items the p value detection method does not have much difficulty in spotting the exposed items. However, it is not realistic to assume that all examinees have prior knowledge to an item but the simulations did inform about the performance of the statistic in this extreme situation.

Generally, hard items are very easy to spot while easy items are hard to spot. This is consistent with our experience. Most examinees are able to answer an easy item correctly without any outside help. While most of examinees benefit more if they have prior knowledge to hard items. If two items have the same difficulties, the one that has high discrimination parameter is easier to spot. Thus, we can conclude from the results that item exposure detection would depend not only on the choice of item exposure detection statistic, sample size, and nature of the exposure, but would also depend on the statistical characteristics of the exposed test items.

The discrimination index also has some impact on the detection process. Generally speaking, higher discriminating items are easier to spot while lower discriminating items are harder to spot.

4.4 Conclusions

When the underlying ability distribution is stable, item p values show apparent invariance over time. This is reasonable since this situation is equivalent to

traditional paper and pencil testing. The estimated item p values are doubtless stable when a big enough sample is picked from the population.

In this study, 100 percent of examinees having prior knowledge is not realistic but it is a valuable starting point for more interesting simulations. Any method should have power in these initial simulations or it would be useful in practice.

In conclusion, item p values are an effective detection statistic when the underlying ability distribution is stable. The advantage of item p value is that it is model free. Not many assumptions have to be made and the user does not have to worry about the fitness of the model or the data. However, it is descriptive rather than statistical. Its primary function is to unveil the trend underlying the sequence. More ingenious tools are needed to account for the trend (if it exists). This first simulation study only provided a base line to understand the concept of moving averages and the usage of the proposed method.

When this method is used in practice, an assumption that the stream of examinees over time is equivalent is implied. However, this assumption may be problematic in some testing programs. For example, examinees with higher abilities may take the test early within a testing window and the poorer examinees may come later, for example. With exams like the GRE, better candidates may tend to test in the fall or winter while poorer candidates may show up in the spring or summer. Besides, item parameter shifts may occur due to changes in curriculum or population characteristics. If this is the case, the trend revealed from the data may result from the different factors that are likely influencing data simultaneously. Even in these situations moving averages are still helpful to uncover the trend underlying the data which may provide desirable information for management purpose. In the next

chapter, some more interesting simulations will be carried out, to further this line of research.

Figure 4.1. Plot of moving p values. (item 01 to item 08, window size=50, ρ = 0)

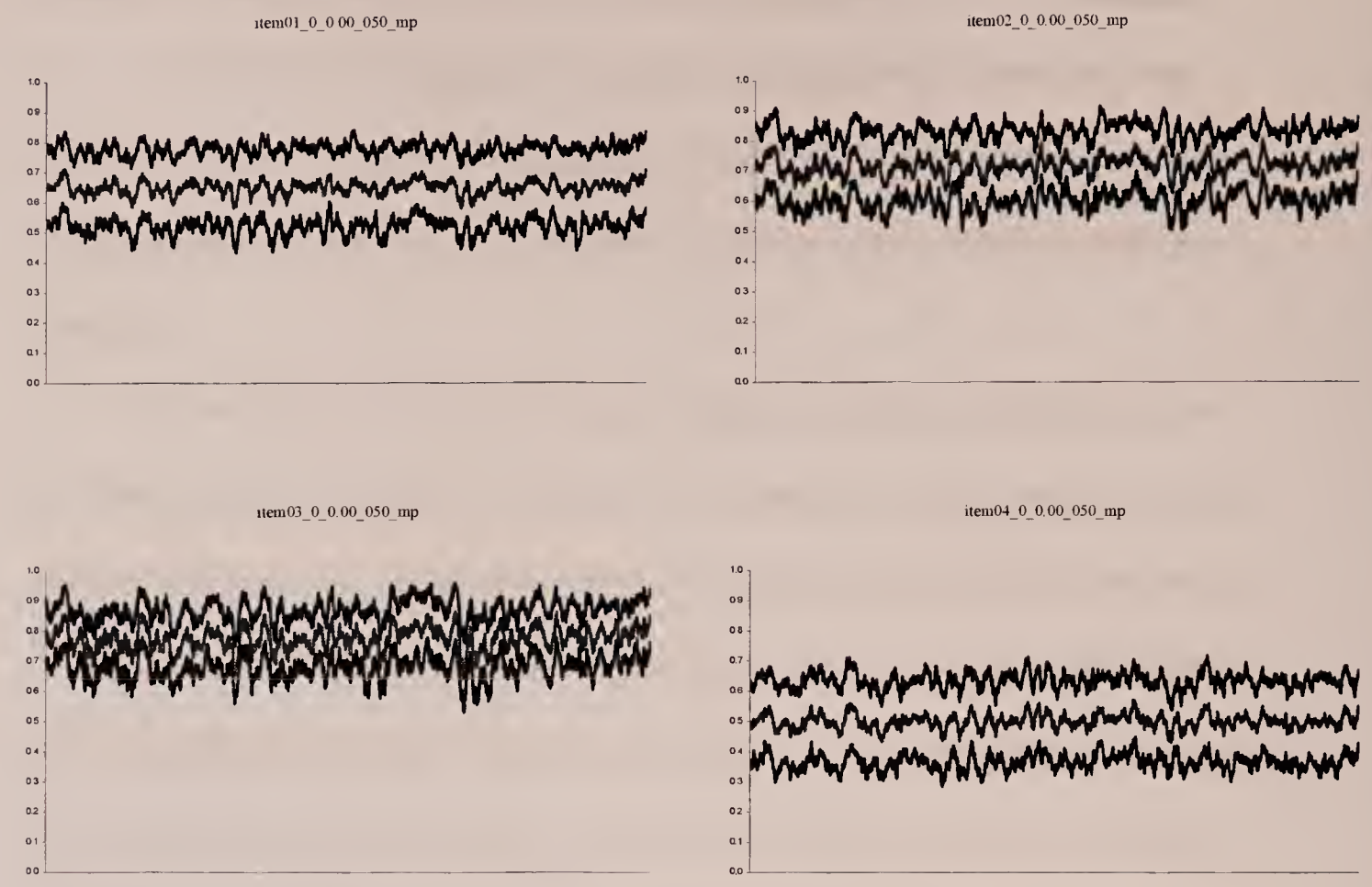


Figure 4.2. Plot of moving p values. (item 09 to item 12, window size=50, ρ = 0)

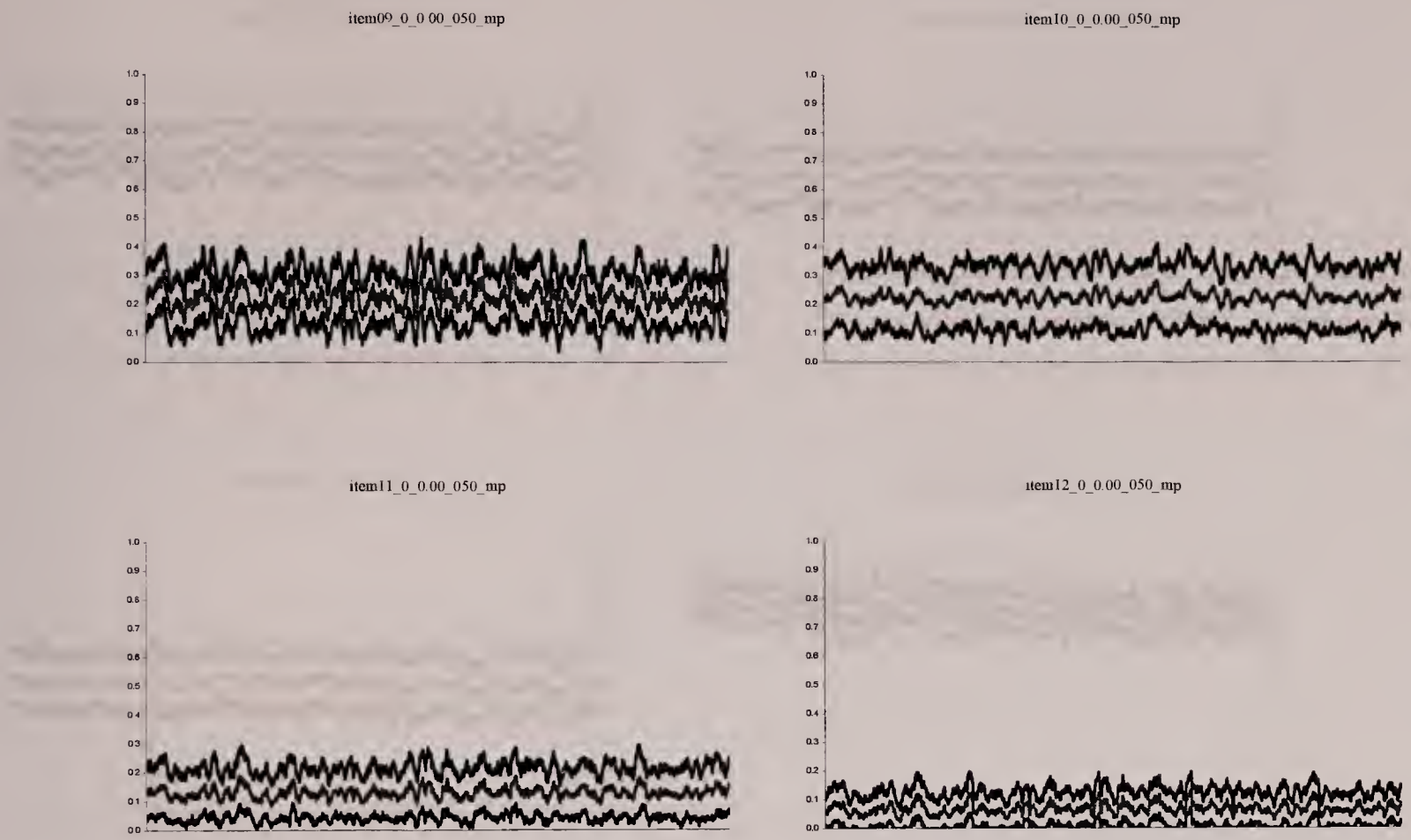


Figure 4.3. Plot of moving p values. (item 01 to item 08, window size=100, ρ = 0)

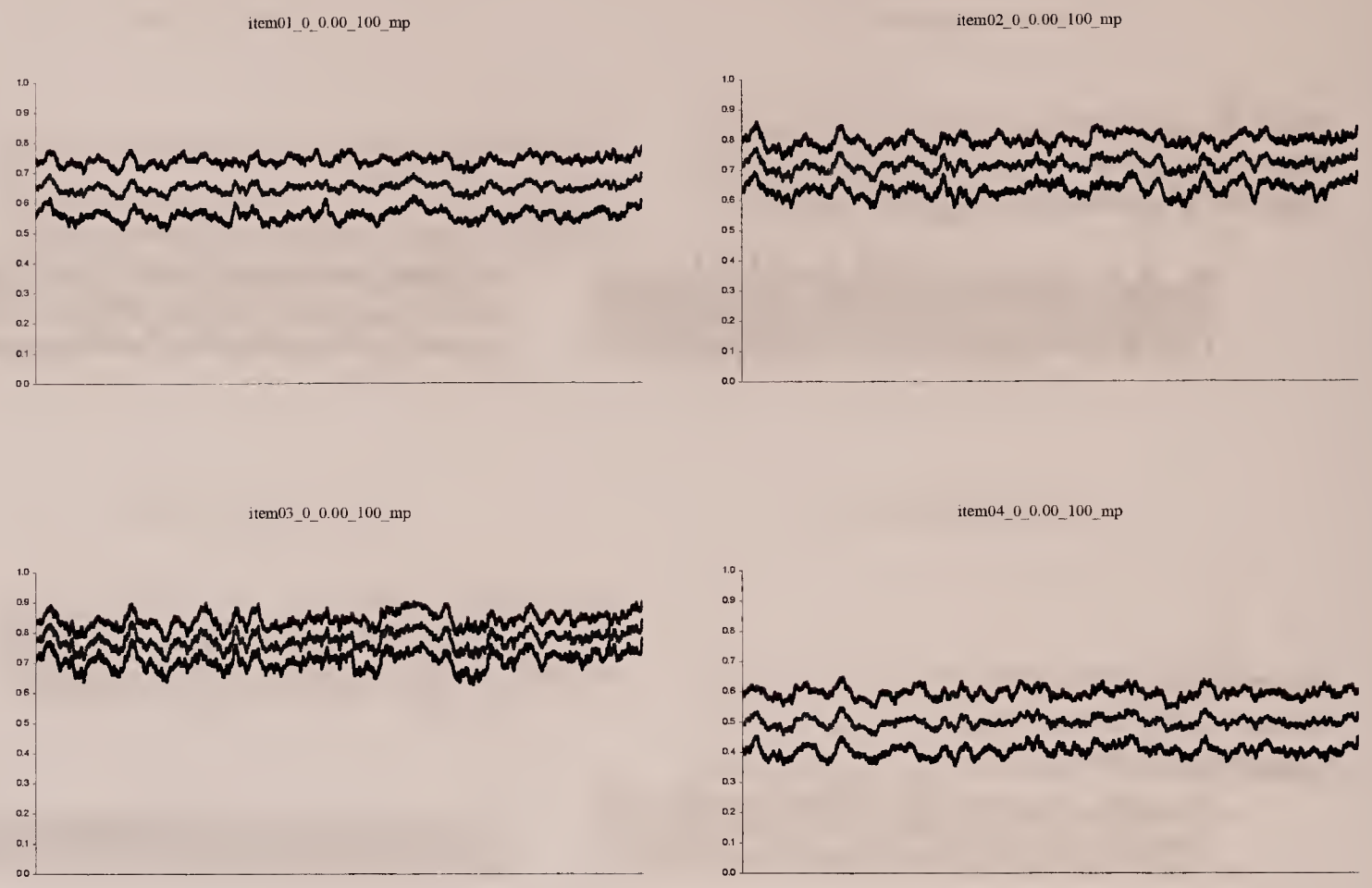


Figure 4.4. Plot of moving p values. (item 09 to item 12, window size=100, ρ = 0)

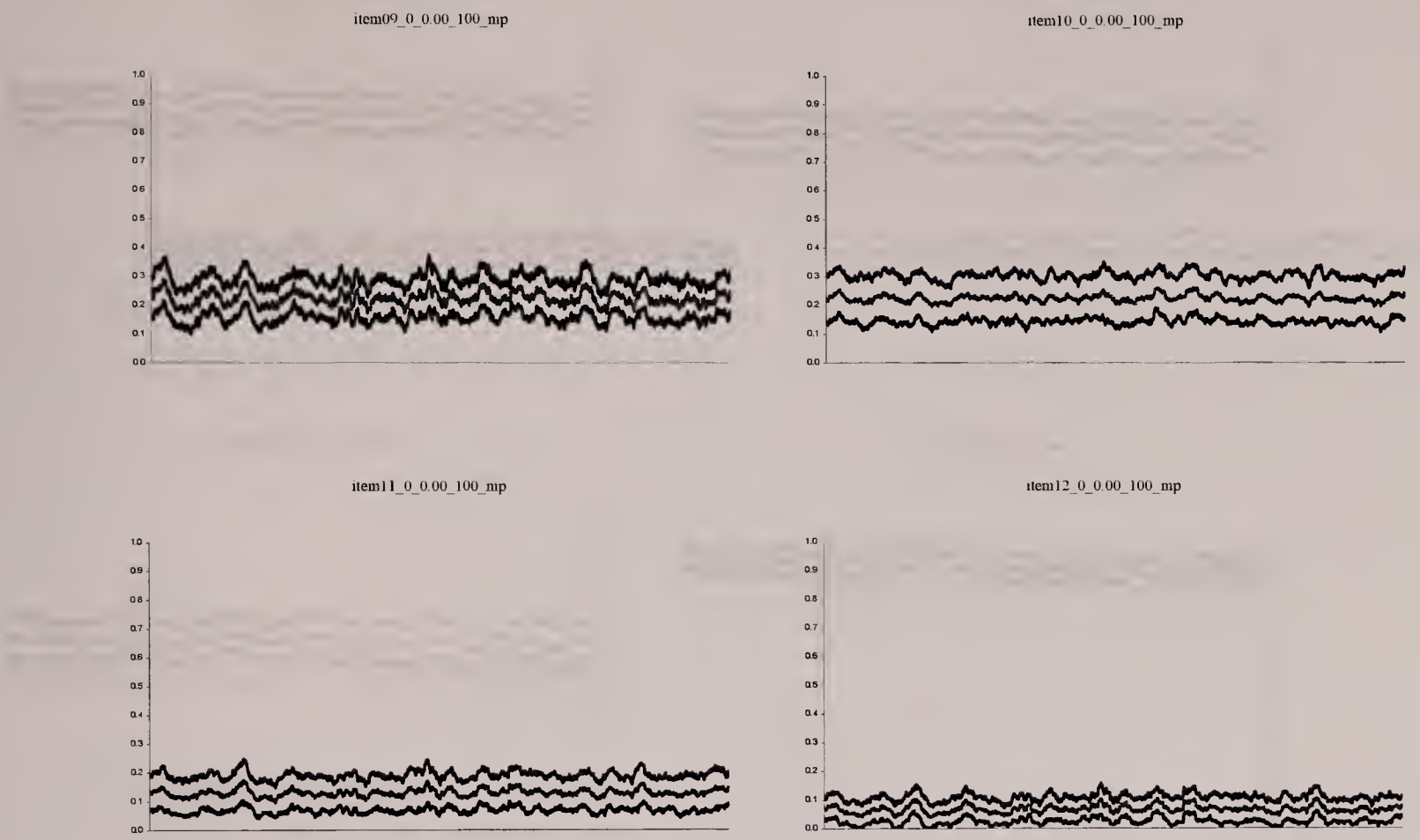


Figure 4.5. Plot of moving p values. (item 01 to item 08, window size=200, ρ = 0)

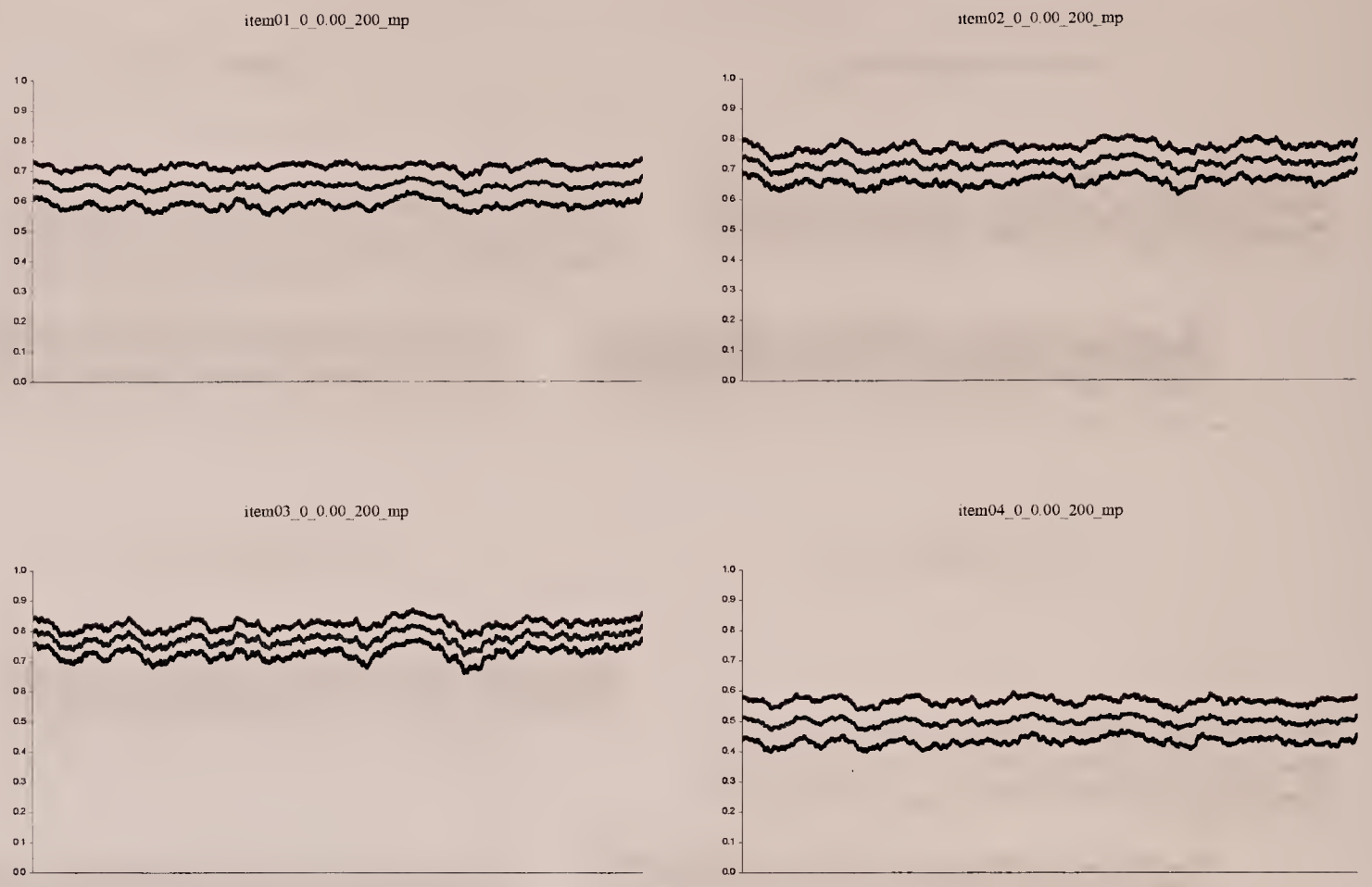


Figure 4.6. Plot of moving p values. (item 09 to item 12, window size=200, ρ = 0)

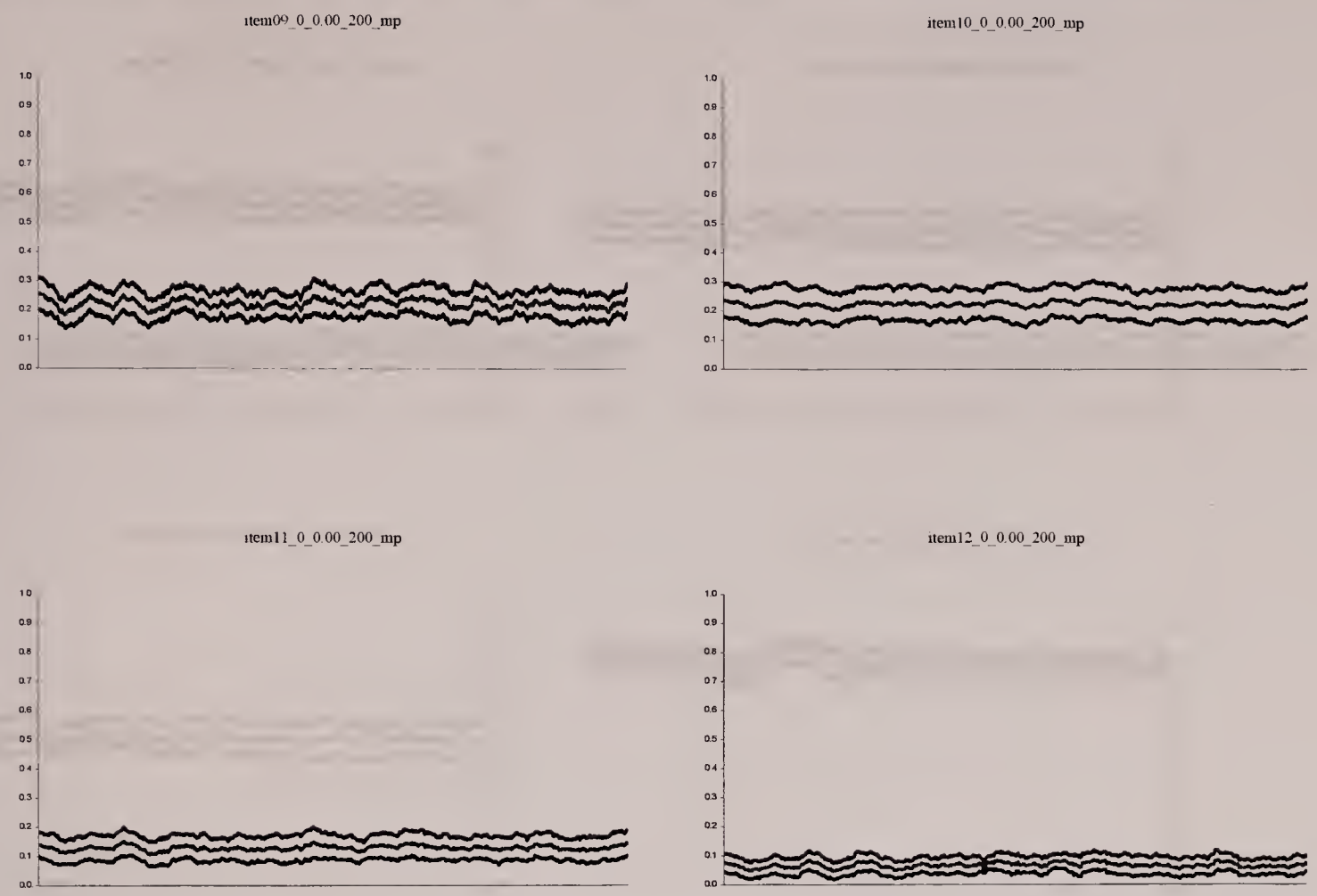


Figure 4.7. Plot of moving p values. (item 01 to item 08, window size=300, ρ = 0)

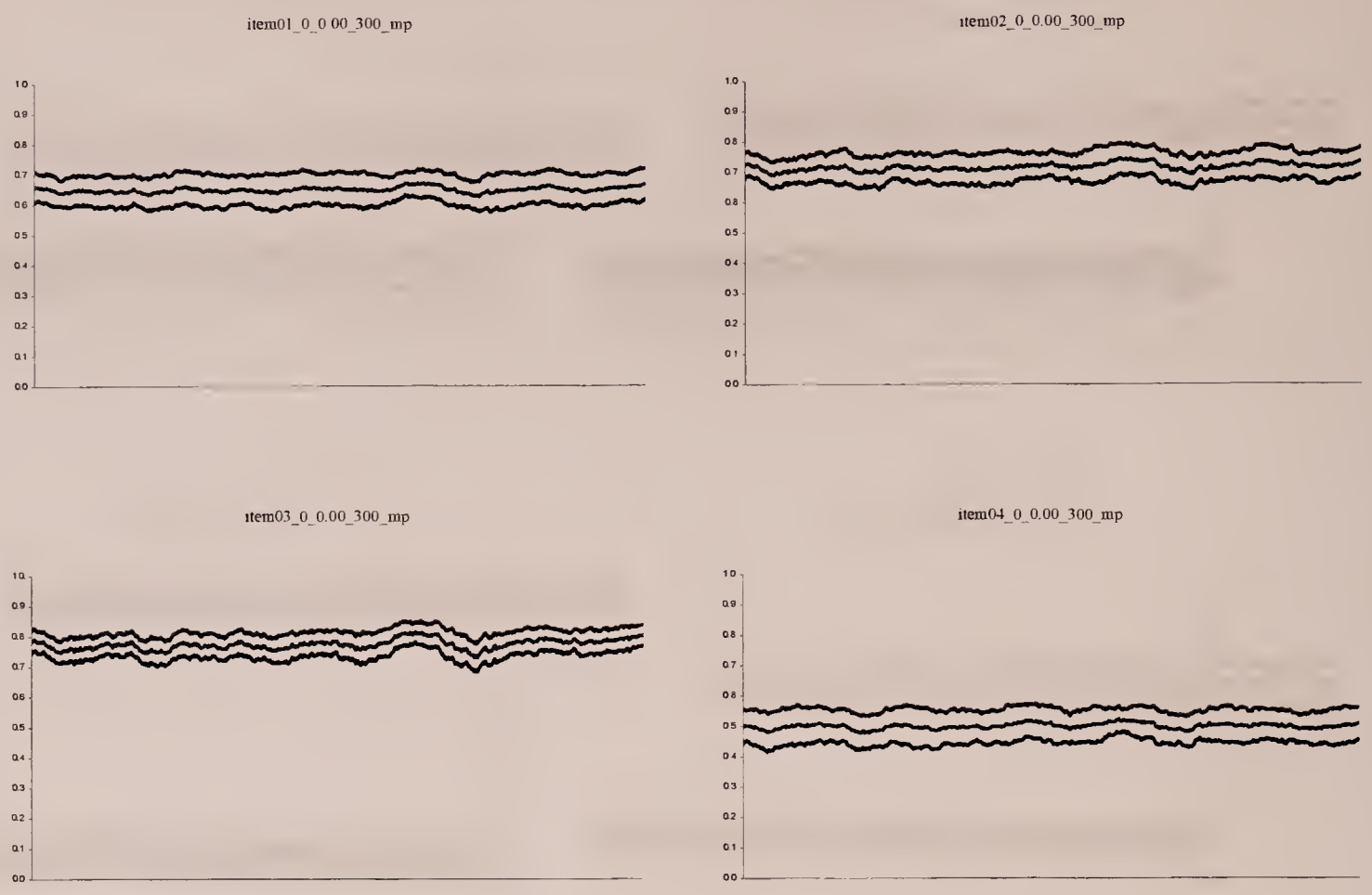


Figure 4.8. Plot of moving p values. (item 09 to item 12, window size=300, ρ = 0)

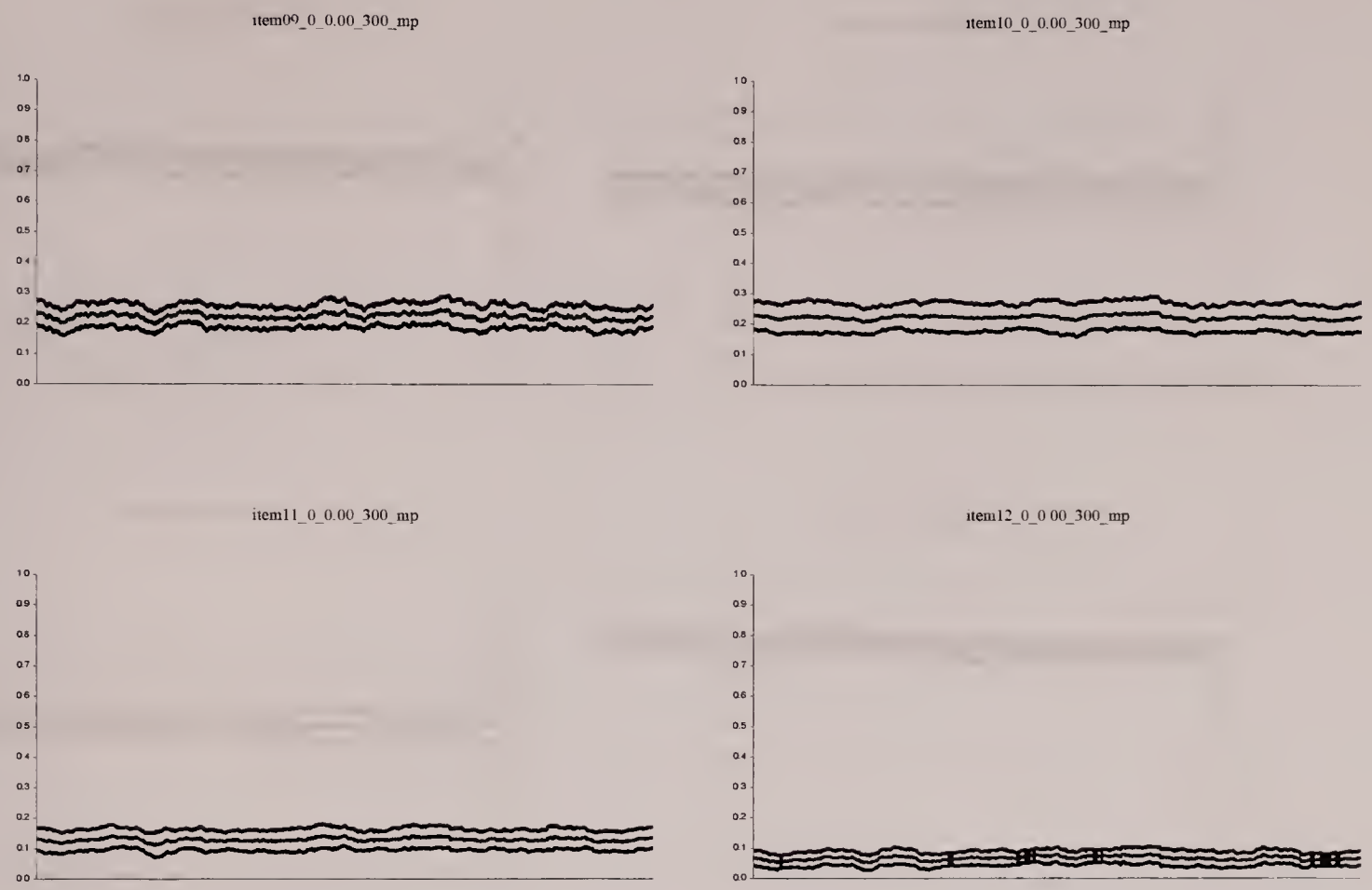


Figure 4.9. Plot of moving p values. (item 01 to item 08, window size=500, ρ = 0)

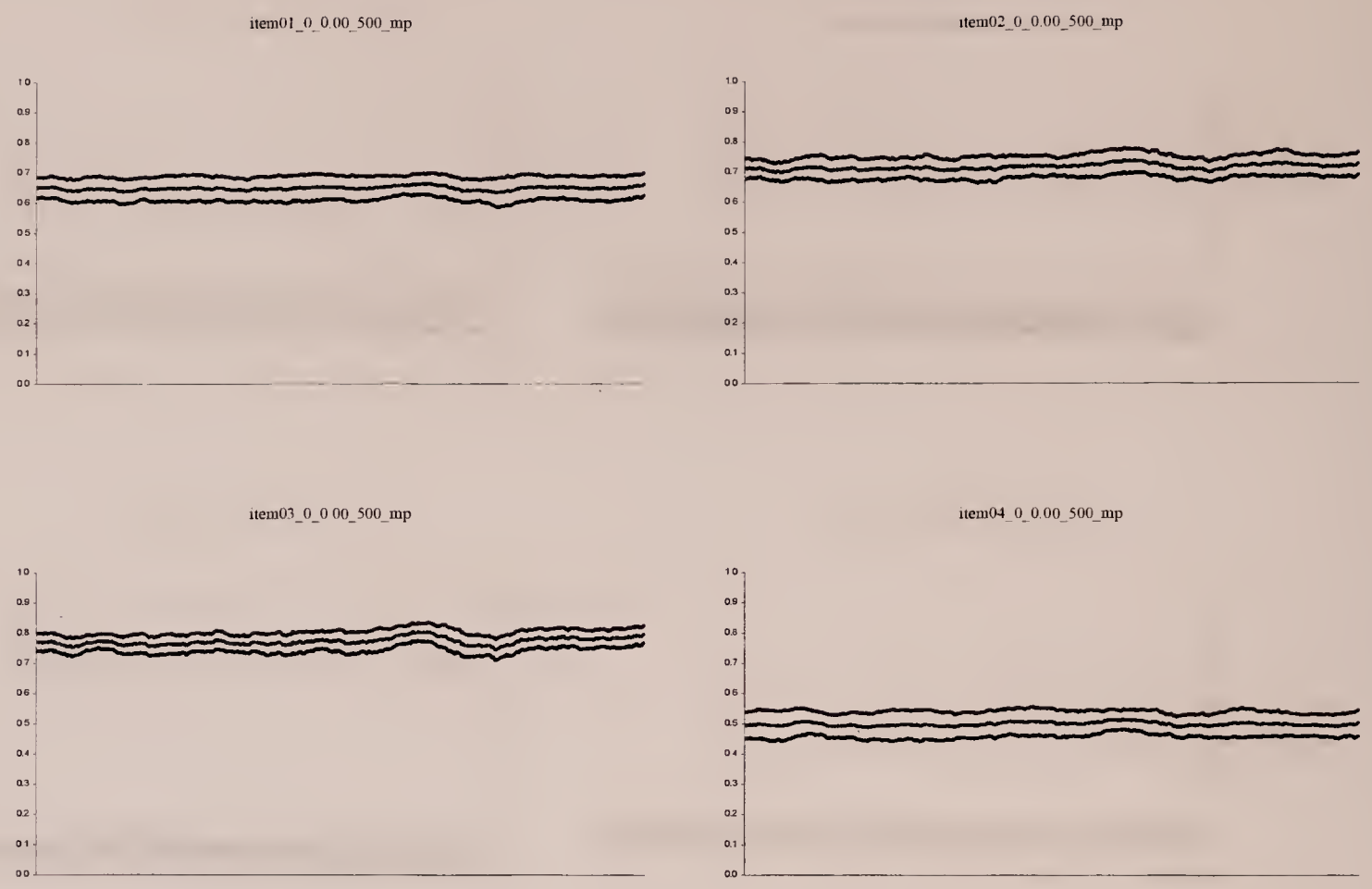


Figure 4.10. Plot of moving p values. (item 09 to item 12, window size=500, ρ = 0)

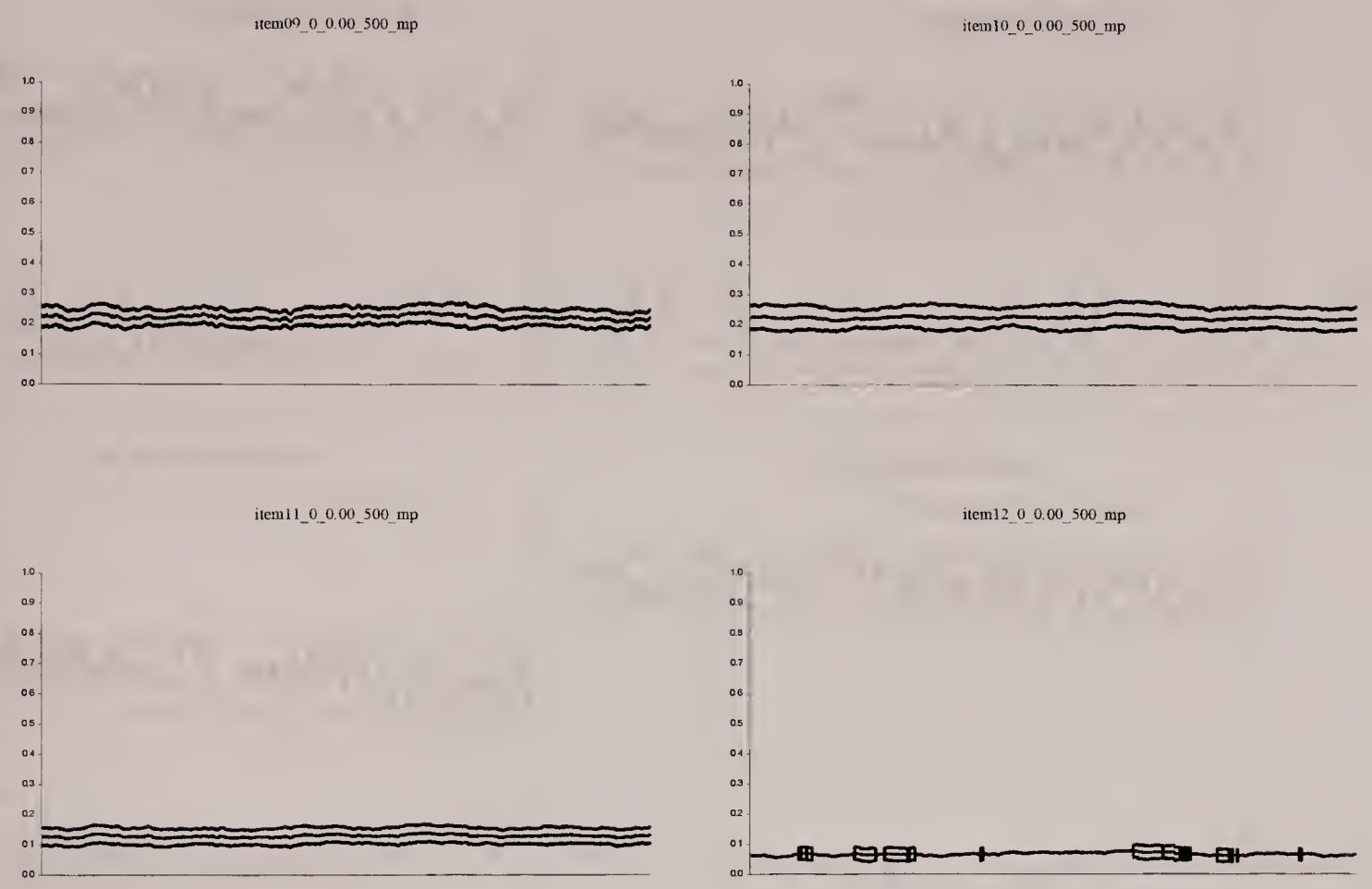
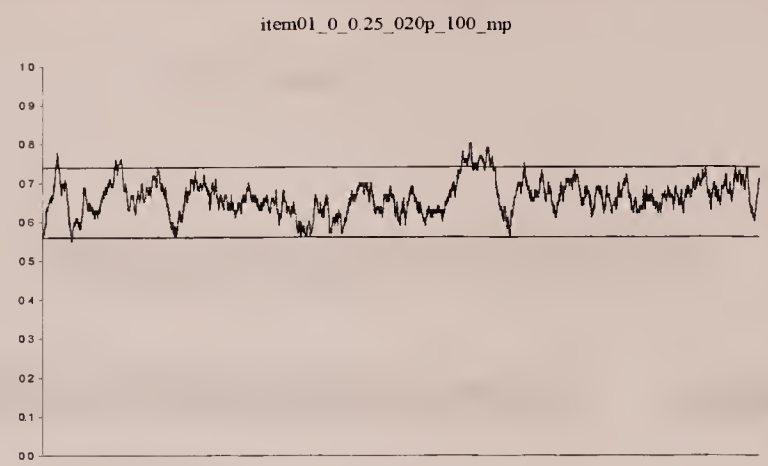
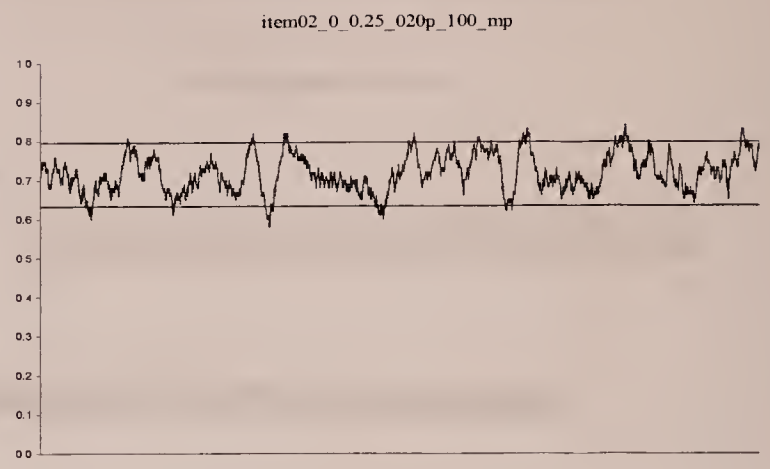
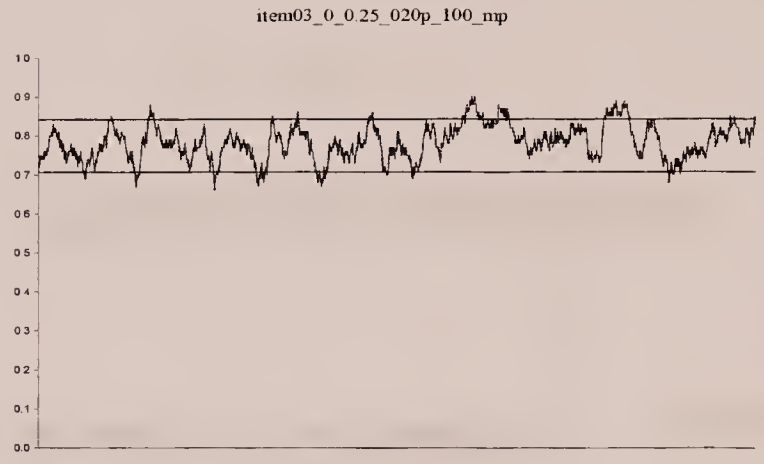
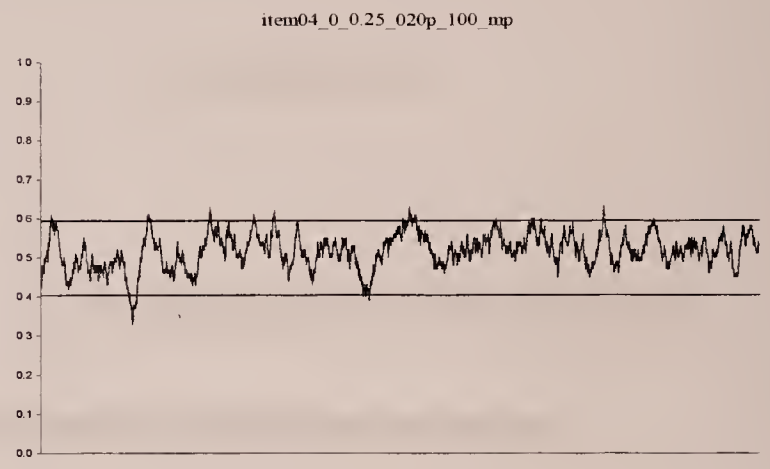


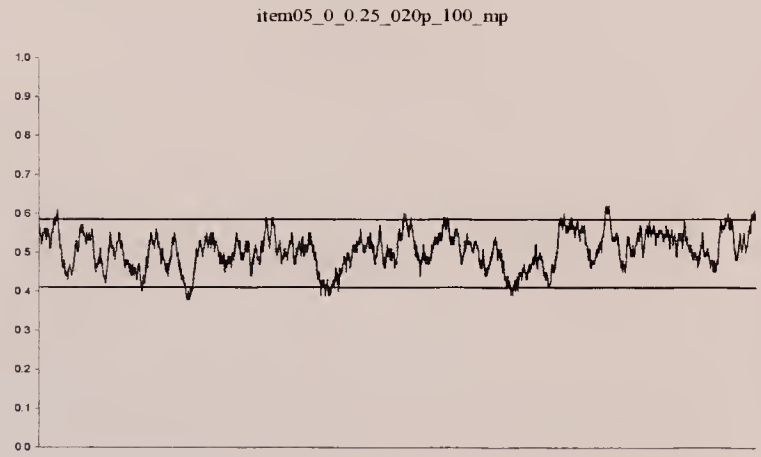
Figure 4.11. Plot of item exposure detecting. (item 01 to item 08, ρ = 0.25, for 20%)

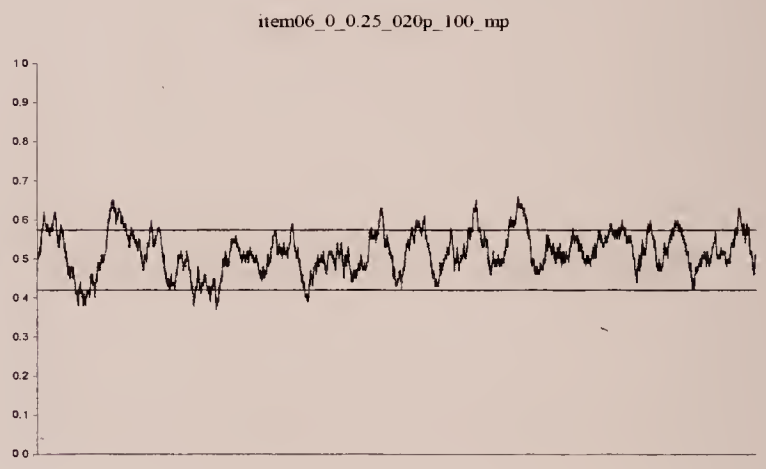


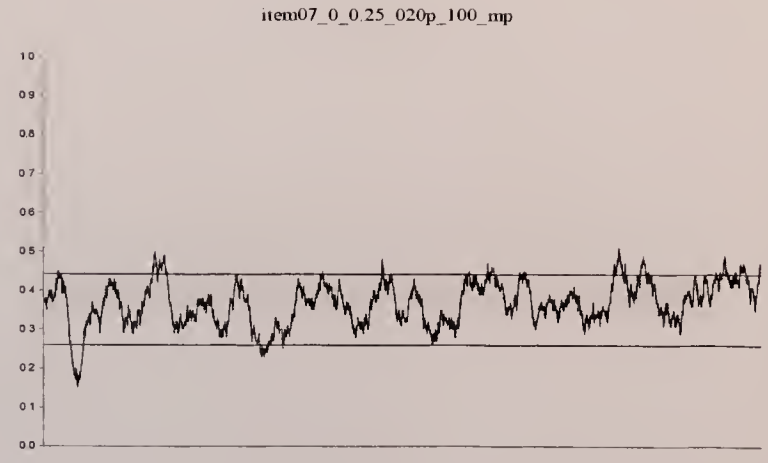












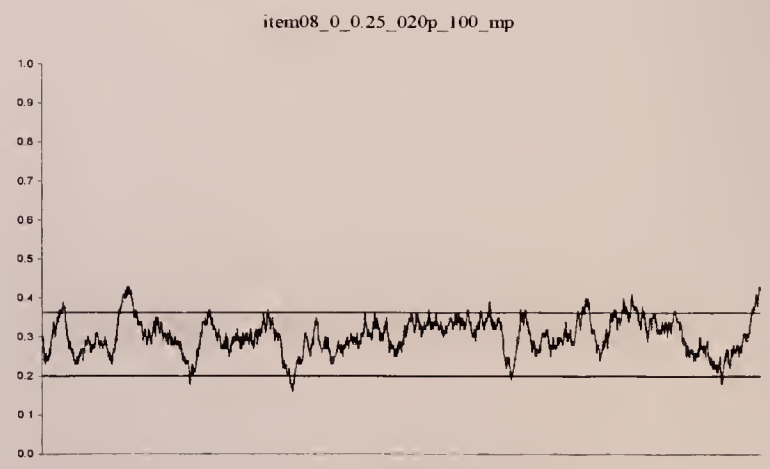


Figure 4.12. Plot of item exposure detecting. (item 09 to item 12, ρ = 0.25, for 20%)

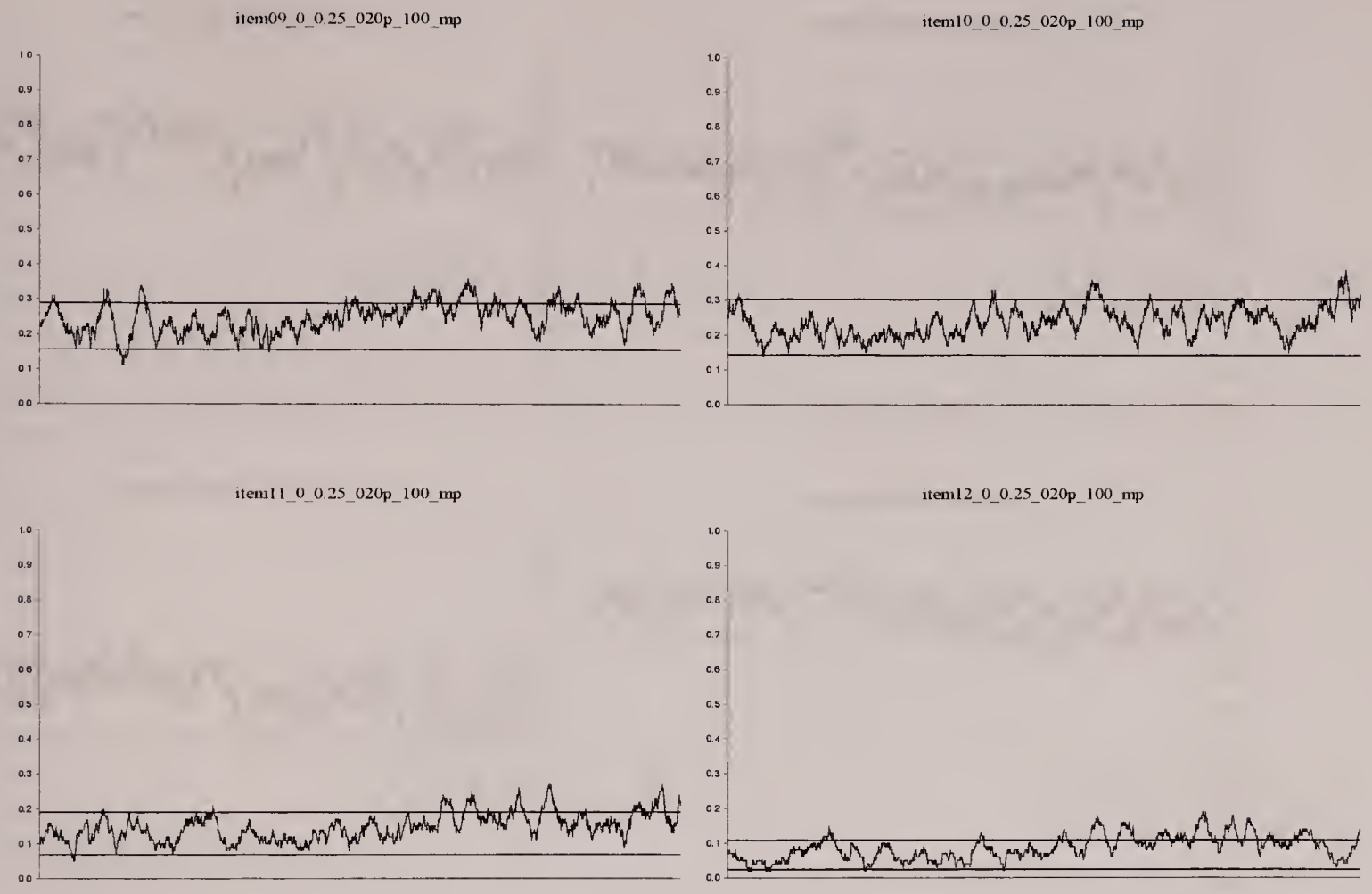


Figure 4.13. Plot of item exposure detecting. (item 01 to item 08, ρ = 0.50, for 20%)

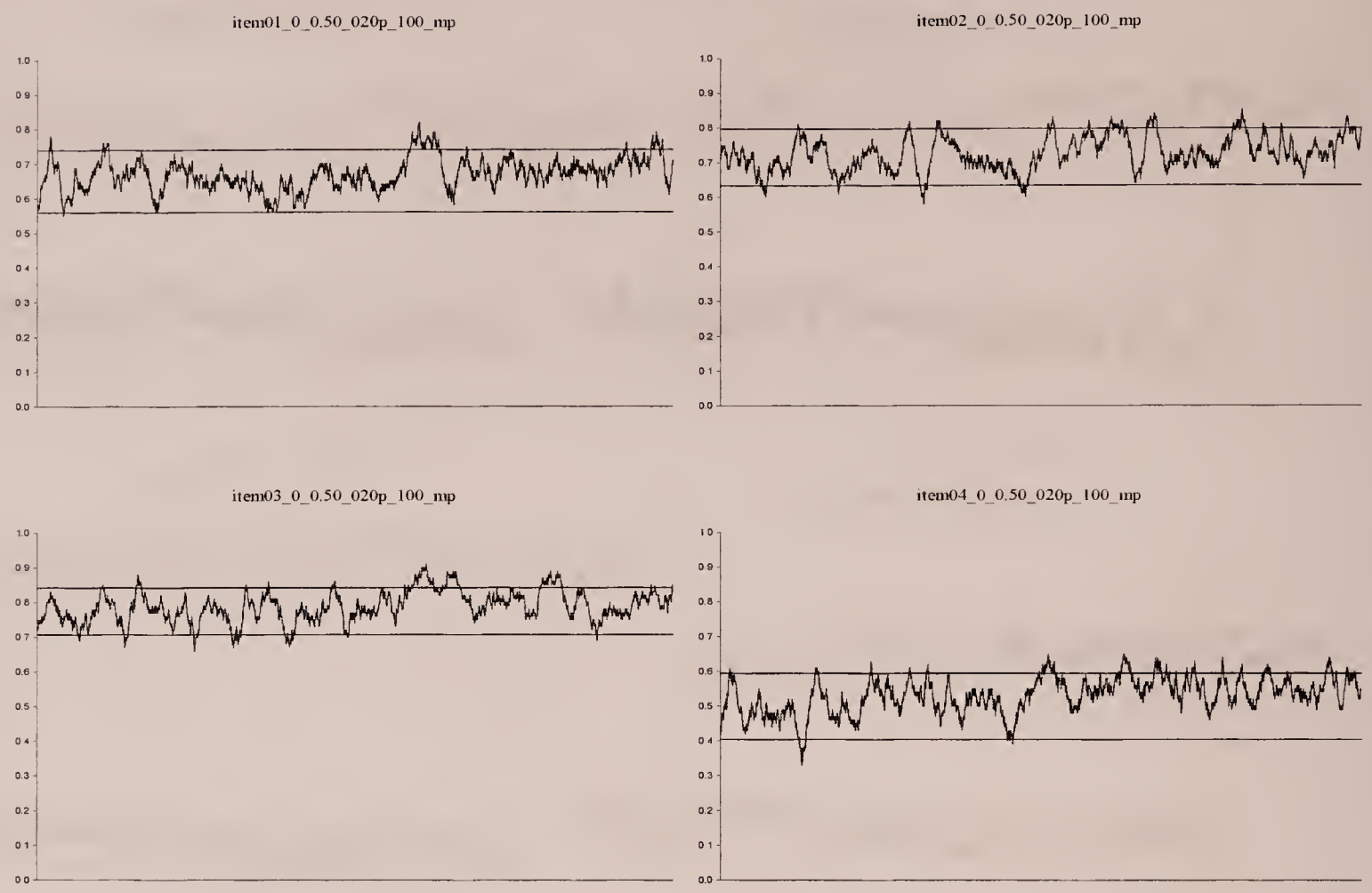


Figure 4.14. Plot of item exposure detecting. (item 09 to item 12, ρ = 0.50, for 20%)

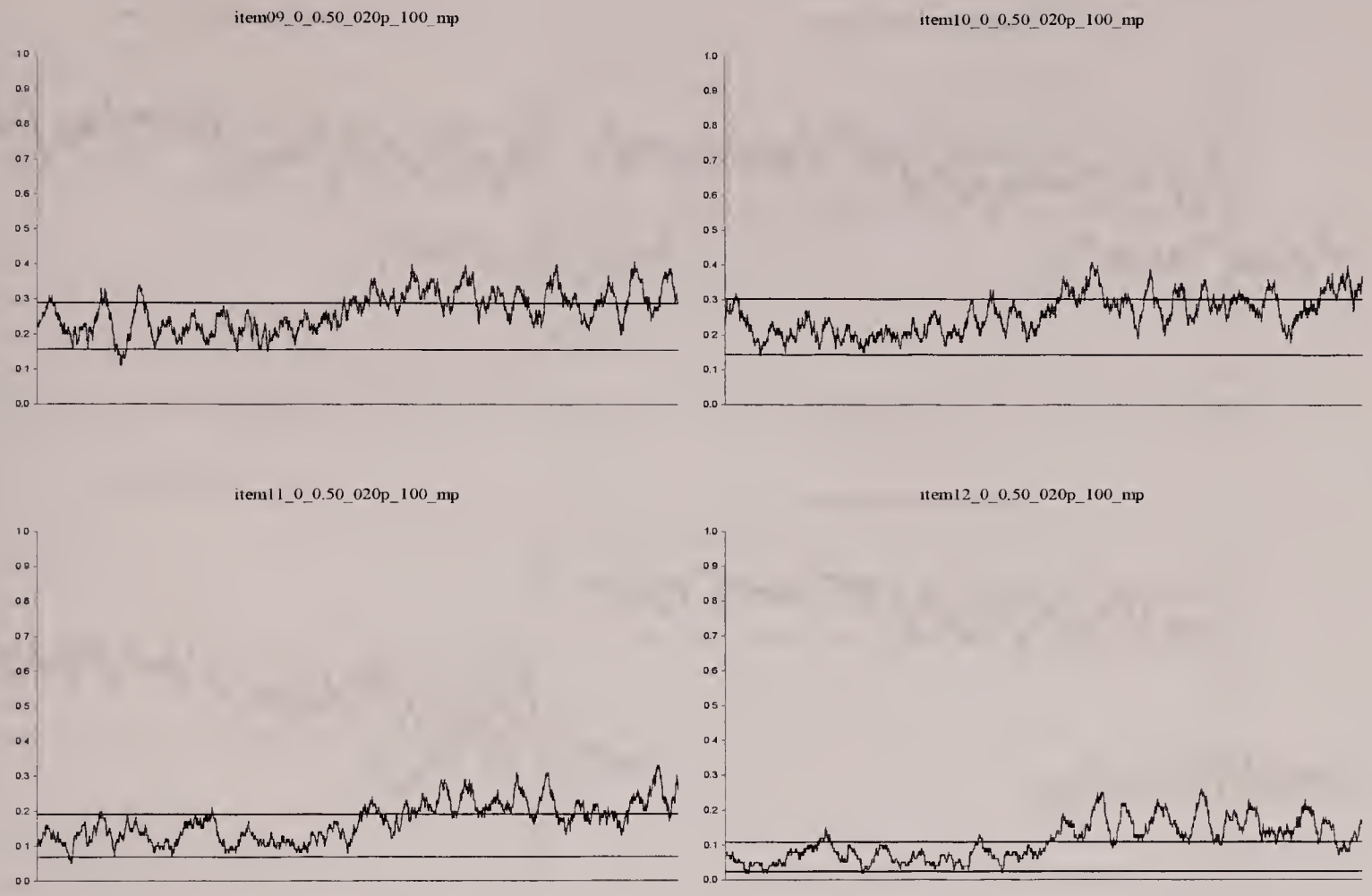


Figure 4.15. Plot of item exposure detecting. (item 01 to item 08, ρ = 0.25, for 100%)

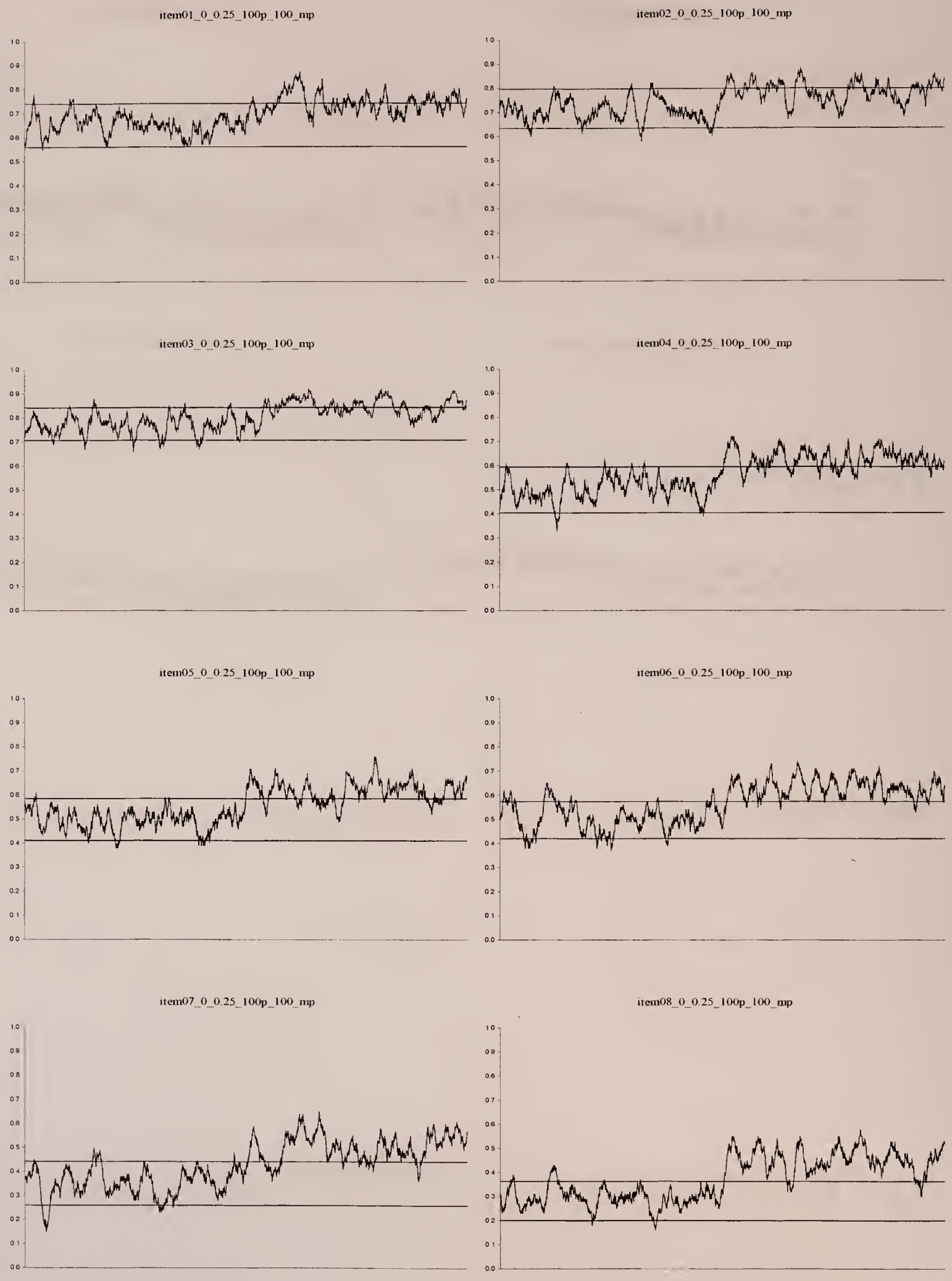


Figure 4.16. Plot of item exposure detecting. (item 09 to item 12, ρ = 0.25, for 100%)

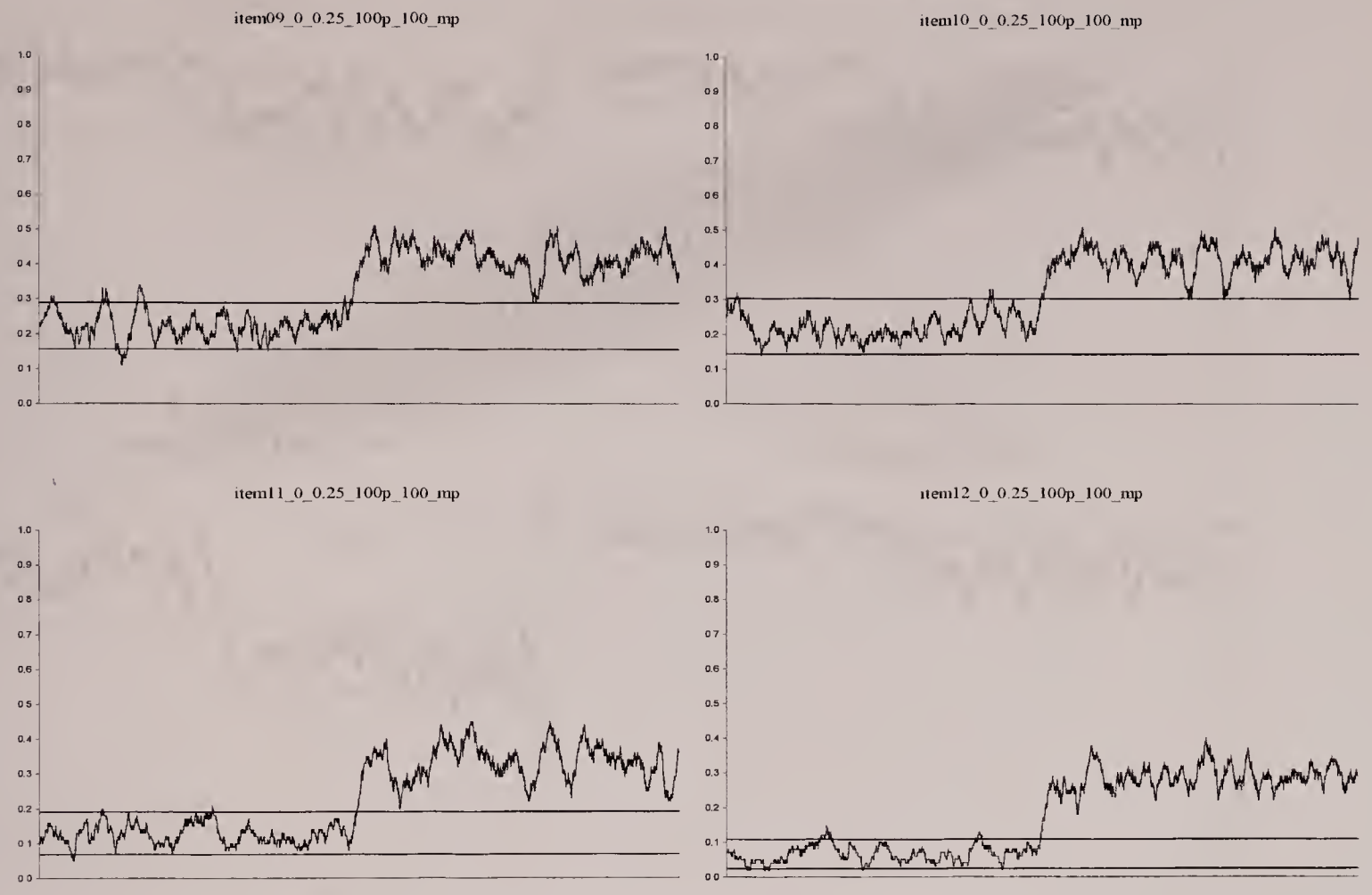


Figure 4.17. Plot of item exposure detecting. (item 01 to item 08, ρ = 0.50, for 100%)

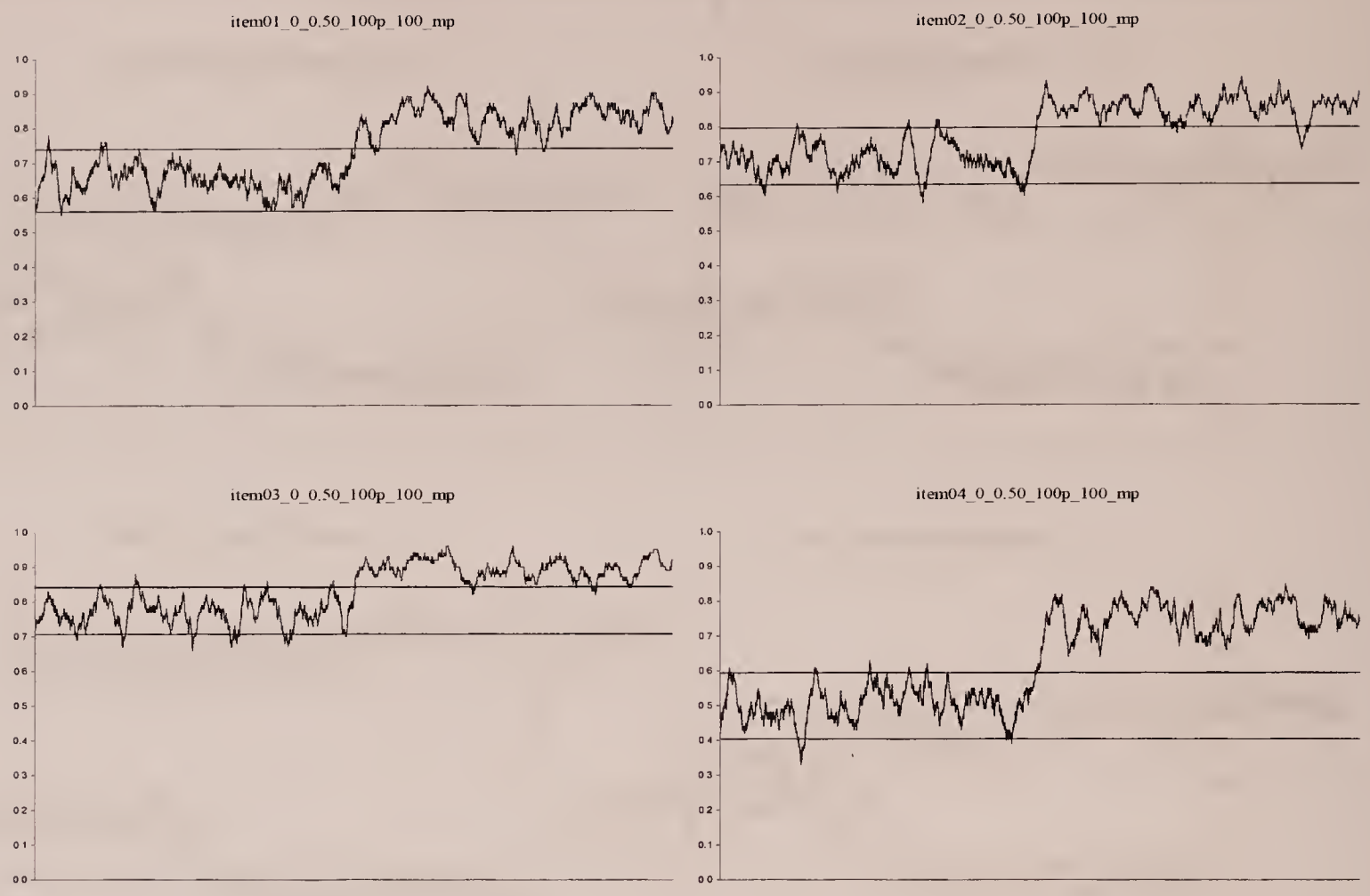


Figure 4.18. Scatter plot of examinees scores ($\rho = 0$ vs. $\rho = 0.25$, for 20%)

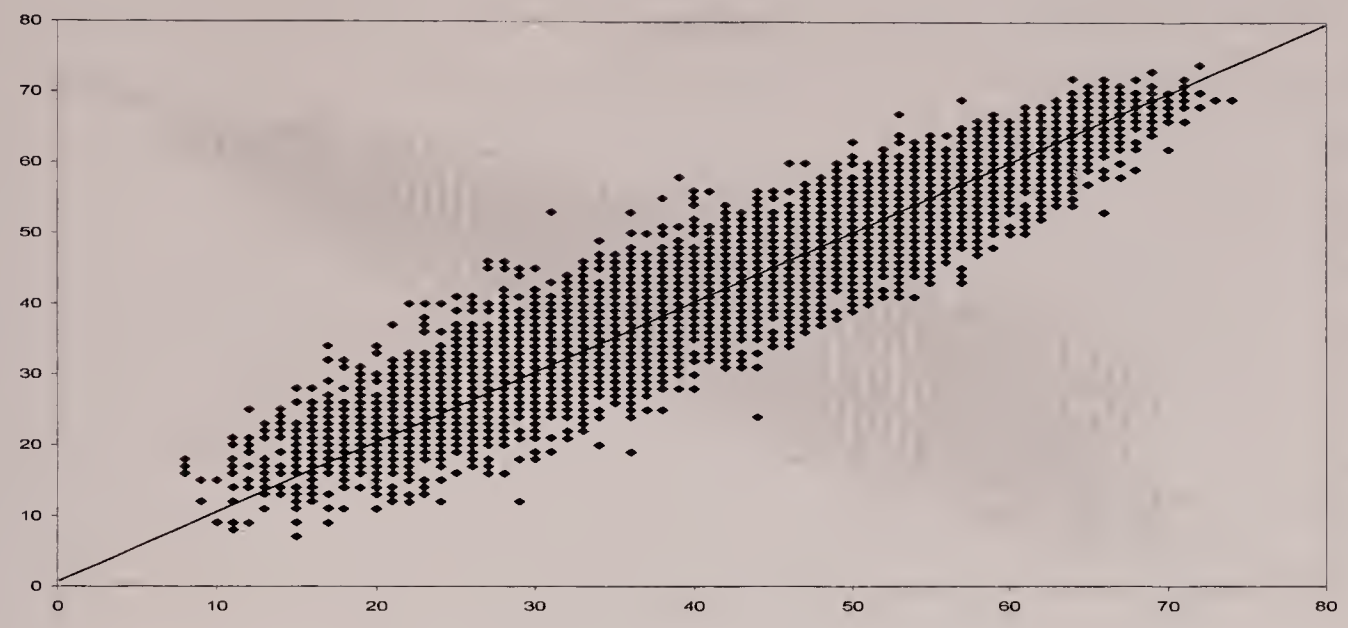


Figure 4.19. Scatter plot of examinees scores ($\rho = 0$ vs. $\rho = 0.25$, for 100%)

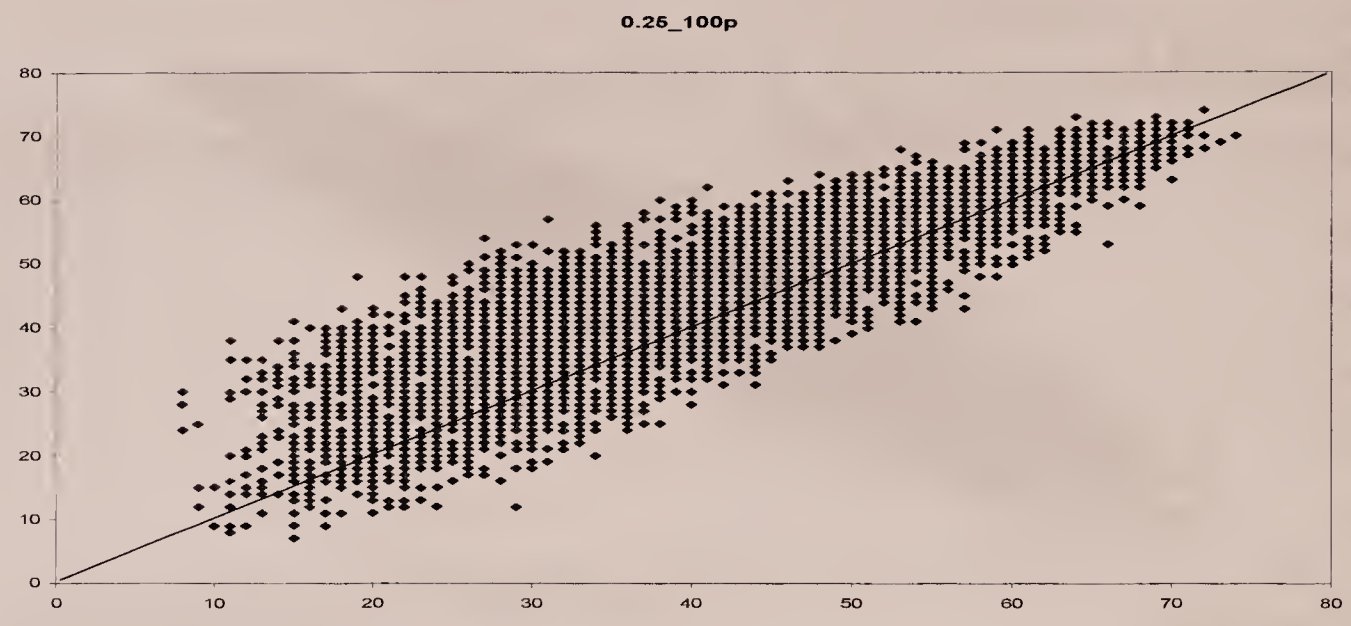


Figure 4.20. Scatter plot of examinees scores ($\rho = 0$ vs. $\rho = 0.50$, for 20%)

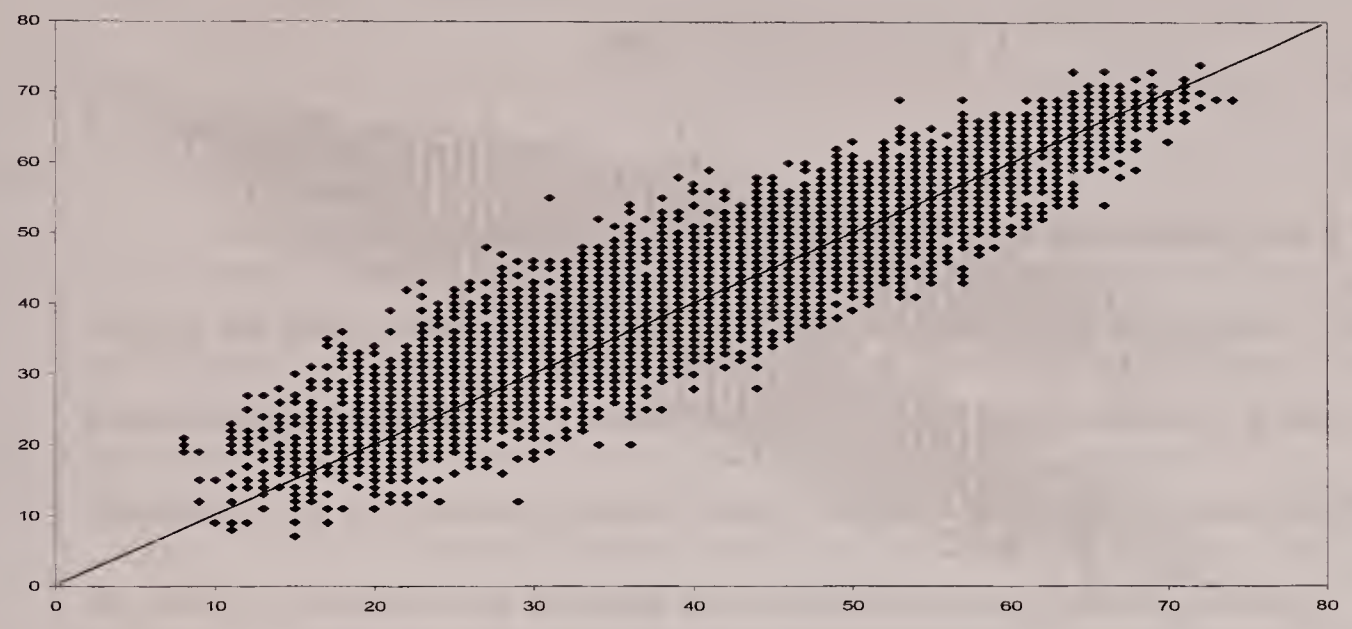
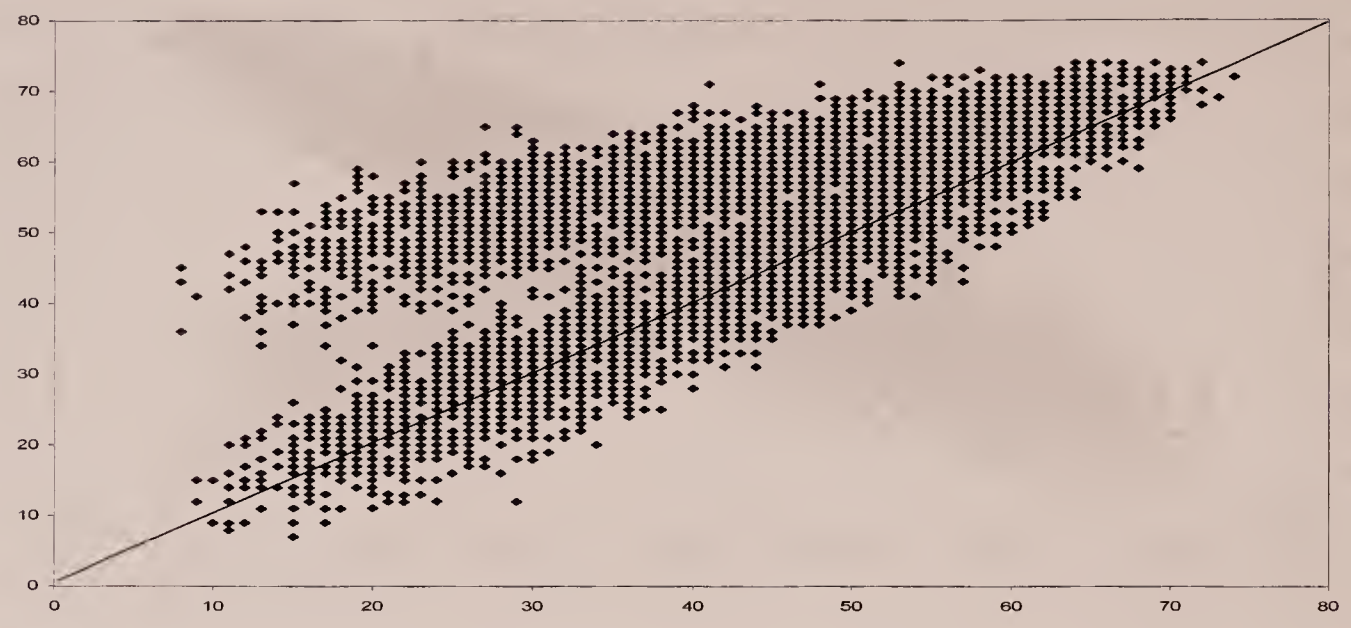


Figure 4.21. Scatter plot of examinees scores ($\rho = 0$ vs. $\rho = 0.50$, for 100%)



CHAPTER 5

SIMULATION STUDY 2

5.1 Purposes

The biggest difference of this simulation study with the previous one is that drifts in the ability distributions were introduced. As discussed in Chapters 3 and 4, the moving p-value statistic is sensitive to ability shifts and therefore, it is less suitable for use by most test agencies doing computer-based testing since shifts in the ability distribution and detection of exposed items using moving p-value averages are confounded. We have to look for item statistics that are free of ability distributions for this case. Thus, in addition to classical item difficulty, two IRT-based item statistics were introduced as detection statistics in this simulation study. As discussed before, these two item statistics were expected to be distribution free because of the property of IRT parameter invariance.

The purposes of this research were (1) to evaluate several item exposure detection statistics in the presence of shifts in the ability distribution over time, (2) to address the suitability of the item exposure detection statistics under a number of item exposure models, and (3) to investigate item exposure detection for items with different statistical characteristics. The first purpose was essential because the assumption that the ability distribution is unchangeable at all times during a testing window is too strong in most educational testing programs. Some drift in the distribution might be expected—for example, the poorer candidates may come first, and higher ability candidates may follow later in the window. Figure 3.8 shows that when there are shifts in the ability distribution the moving p value sequence will show the same trend with that of ability distribution. Even though sometimes it may be true that the ability distribution of candidates will by-and-large be equivalent over

time, item exposure detection statistics that are free of this questionable assumption should be studied. As shown in Chapter 3, IRT-based statistics are promising.

Achieving the second purpose would provide data on competing item exposure detection statistics under various item exposure models. For example, in one simple model, after an item is exposed by a candidate one might conjecture that all candidates will have knowledge of the item and answer it correctly if it is selected for administration again. Several other item exposure models need to be investigated too, several that are a bit more realistic.

The third purpose was important because the results of Chapter 4 had shown that the item exposure detection rate would depend not only on the choice of item exposure detection statistic, sample size, and nature of the exposure, but would also depend on the statistical characteristics of the exposed test items. For example, items 1 and 2 were very difficult to detect since they were easy for candidates. Most of the candidates were already expected to do well and any improvements in item performance due to exposure then would be small. Harder items should be considerably easier to spot because the shifts in item performance due to exposure are likely to be greater.

5.2 Details of the Methodology

Variables under study included (1) ability distribution (fixed or variable), (2) choice of item exposure detection statistic, (3) type of item exposure model, and (4) statistical characteristics of exposed test items.

At the first step of the present study, the level of item exposure controlling parameter ρ was varied from no exposure ($\rho=0$) to full exposure ($\rho=1$). The situation of no exposure ($\rho=0$) shows whether or not the item statistics were free of ability distribution while full exposure ($\rho=1$) tells us how the item statistics perform during

the extreme situations. For non-zero item exposure controlling ρ, two levels (10% or 100%) of examinees have prior knowledge to the exposed items were simulated. Again 100% was interesting since we wanted to know the extreme situation but 10% was a more realistic situation.

The situation of no exposure (ρ=0) is important since these simulations set up a base line for the value distribution of moving averages. This information was used to set up the control limits. This is especially useful for item residuals. Although there is no strict proof, there is lots of empirical evidence showing that the standardized residuals have an approximate normal distribution (see, for example, Hambleton, Swaminathan, & Rogers, 1991). However, the control limits for all item statistics were all set up by the simulation study for the situation of no exposure (ρ=0).

An intermediate value of ρ=.25 applied to either 10% or 100% of the candidates was then considered in the simulations. ρ=.25 and 10% was considered as a minimum noticeable exposure according to the simulation study we reported on in the previous chapter. We expected the proposed item statistics would perform well in practice if they performed well in this situation.

Three different ability distributions for the 5000 candidates were considered: The first one was a normal distribution with mean of zero and standard deviation of one. The second one drifted from a lesser ability group to a higher ability group. To be specific, when generating the *i*th examinee, proficiency or ability was randomly generated from a normal distribution $N(-1+\frac{i}{2500},1)$. That is to say, the first examinee was generated from N(-0.9996, 1), which is very likely to be a small number. On the other hand, the 5000th examinee was generated from N(1,1), which is very likely to be a bigger number than for previous examinees. By this simulation

approach, we were assuming that the poorer candidates, generally, would take the test early (average ability = -1.0) and then gradually the ability distribution would shift from a mean of -1.0 to a mean of +1.0 by the end of the testing window. The third one shifted abruptly from a lower ability group to a higher ability group. The first 2500 candidate abilities were sampled from a N(-1.0, 1). For the last 2500 candidates, candidate abilities were sampled from a N(+1.0, 1) distribution.

In Chapter 4, an item detecting chart was used to show how item exposure can be displayed visually. In this chapter, more accurate indices were computed to indicate the properties of the statistics, which are power and type I error rate. Power and type I error rate were computed empirically by the method described in Chapter 3, Section 3.6.

To compute the power and type I error, the simulation study was replicated 100 times while the plots displayed a single simulation process.

5.3 Findings

Two types of tables were reported: The first type of tables inform how many times an exposed item had been administered from the time point when the item was exposed to when it was detected. The second type of tables inform about the power and type I error for each statistic. There are two columns for each item in this type of table. One labeled with "I" indicates that among the first sub-sequence of the examinees (500^{st} to 1549^{th}), the proportion of the moving average sequence that will exceed the upper limit. The other one labeled with "II" indicates that among the second sub-sequence of the examinees (3500^{st} to 4499^{th}), the proportion of the moving average sequence will exceed the upper limit. Since the item was not exposed for the first sub-sequence and exposed for the second sub-sequence, the

figures under label "I" provided information of type I error rate while the figures under label "II" provided information of power.

Figures 5.1 highlights the functioning of the three item statistics for a medium difficult item (b=0.0, a=0.7) with normal ability distributions while Figure 5.2 to 5.3 highlight the functioning of the two IRT based item statistics for the item with shifting and abrupt change in ability distribution, respectively. With a fixed normal distribution, all three item exposure detection statistics are quite stable as they should be. With a shift in the ability distribution, gradual or abrupt, we have learnt that the p-value statistic shifted as well, and substantially so (see Figure 3.7). Clearly, in this case p-value shifts are confounded with shifts in ability distributions and not reflecting item exposure because there was no exposure. So the moving p value was dropped from the other two plots. Obviously it is not interesting at all to look at the trend that consists of information both about abilities and item exposure. The two IRT-based item exposure statistics, as we expected showed excellent stability on the ability continuum. This finding is not surprising at all. As long as the IRT model fits the data well enough, the sum of the residuals for a group of examinee should be very close to zero and with an expected value of zero. An interesting point is that when item exposure exists, the moving averages seem to be bigger at the lower ability end and smaller at the higher ability end.

What we mean when we say these two IRT-based statistics are free of ability distribution is that when there is no item exposure the moving averages should be around zero. When item exposure exists, the values of the moving averages of these two indices do depend on the abilities of examinees. The moving averages will be bigger for lower ability examinees and smaller for higher ability examinees. This is reasonable since higher ability examinees benefit from item exposure less than

lower ability examinees do. It is very likely that it is faster to spot an item if more lower ability examinees have prior knowledge to the exposed item.

Tables 5.1 to 5.24 contain the relevant information about speed of detection, type I error rates and power of detection, for items with various statistical properties under different item exposure parameters.

Tables 5.1 to 5.8 provide the results obtained with a constant normal distribution of candidate ability. Here, all three item exposure detection statistics were expected to be potentially useful and they were. Table 5.1 shows that with ρ=1.0, with 100% of the examinees benefiting from the exposed information on the 12 items, that detection was very fast. Across 100 replications for example, Table 5.1 highlights that with b=-1.00 and a=0.40, the average number of examinees who saw the exposed item was 27.4 before the statistic exceeded the threshold. (Note that in the simulations, exposure always occurred with the 2501st student in the sequence of 5000 candidates who would see the item.) Detection was even faster with harder items. And, in general, more discriminating items were detected faster too, except when the items were on the easy side. There were very little, if any, differences among the item exposure detection statistics. They all functioned about the same and functioned well.

Table 5.2 shows the type I and power statistics for the 12 items. Type I errors were based on data compiled from the 1500^{th} administration of the item to the 2500^{th} administration. In this portion of the window, there was no item exposure. It is seen in Table 5.1, that under the conditions simulated, the type I error rate varied from 1.5% to 2.7% with the low discriminating items and was somewhat closer to the 5% level with the more discriminating items (2.6% to 4.4% with a=.7, and 1.9 to 6.6% with a=1.2) which had been the goal. More important, was the level of power

of detection. In the case with ρ=1.0 and 100% exposure, detection was very easy and the power of detection was 100% for all items. Figure 5.4 shows what was going on graphically with a normal distribution of candidate ability.

Table 5.3 presents the first set of interesting results for the case where only 10% of the candidates have exposure to the item. Again, the more difficult items are spotted after considerably less item administrations that than easier items. For example, with b= -1.0, a=0.40, 320.7 (on the average) candidates were administered the easy item prior to exposure being detected with the moving p value item exposure statistic. With the hardest item (b=+2.0), and with the same item exposure detection statistic, 98.5 (on the average) candidates were administered the item prior to exposure being detected. With the other item exposure statistics, exposure appeared to be a bit quicker. In general, more discriminating items were detected faster than less discriminating items if they were medium to high difficulty.

Table 5.4 shows, for example, that type I errors were in the 1.5% to 6.6% range across all of the combinations of runs. Choice of item exposure detection statistic was of no major significance in the findings. Perhaps the most noticeable result in Table 5.4 is the low power of detection of exposed easy items (b=-1.0 or b=0.0). 25.2% detection rate was the highest. Whereas for the more difficult items (b=1.0 and b=2.0), power of detecting exposure ran as high as 94.7%. Clearly too, for the more difficult items, detection rates were higher for the more discriminating items. For example, considering the most difficult item (b=2.0), with the standardized item residual statistic, the power rates for items with discrimination levels of .4, .7, and 1.2, were 49.4%, 74.9%, and 93.5%.

Table 5.5 presents the first set of results for the case where ρ=0.25 and 100% of the candidates had exposure to the 12 items. Detection of item exposure did not

take very long. Here again, the more difficult items were spotted after considerably less administrations than easier items. For example, with a=0.40, 113.5 (on the average) candidates were administered the easy item (b=-1.0) prior to exposure being detected with the moving p value item exposure detection statistic. With the hardest item (b=+2.0), and with the same item exposure statistic, 39.5 (on the average) candidates were administered the item prior to exposure being detected. With the other item exposure detection statistics, detection of exposure appeared to be a bit quicker, but only marginally. In general, more discriminating items were detected faster than less discriminating items if they were medium to high difficulty.

Table 5.6 shows, for example, that type I errors were in the 1.5% to 6.6% range as noted before across all of the combinations of runs. Choice of item exposure detection statistic was of no major significance though the two IRT-based statistics appeared to function a bit better overall. This time, detection rates for exposed easy items ran about 35 to 40%, compared to a detection rate of 100% for the hardest items.

Table 5.7 presents the poorest detection rates of the four item exposure models (ρ=.25, 10% exposure). Even for the most difficult and discriminating items, nearly 200 administrations were needed. In the main though, trends were the same: More difficulty and more discriminating items took less time to detect than the easier items. In this condition, interestingly, the moving p value item exposure detection statistic actually functioned a bit better than the other two statistics. Residuals and standardized residuals cross each other on different items. This reminds us that classical item difficulty shows some potential when it is reasonable to assume that the ability distribution is stable over time.

Table 5.8 shows that the likelihood of detecting exposure was very poor. Even for the most difficult and discriminating items, power of detection did not exceed 26%. Choice of item exposure detection statistic was of no major significance.

Figures 5.4 to 5.7 highlight the pattern of the item exposure detection statistics for item 5 (b=0.0, a=0.7) under the four item exposure models with a normal distribution of ability. What is seen is the following: For ρ=1, and 100% exposure, the item was very easy to detect (see Figure 5.4); for ρ= 0.25, 100% exposure, the item took somewhat longer to identify and the power was moderate (see Figure 5.6); for ρ=1.0, 10% exposure, the trend was clear but the item was not identified very often (Figure 5.5); and finally with ρ=.25, and 10% exposure, the exposure was barely detectable in the moving average lines. These figures were presented for illustrative purposes only, and for accurate information on power of detection associated with specific items, see Tables 5.1 to 5.8.

Tables 5.9 to 5.16 and Figures 5.8 to 5.11 contain the statistical results for the gradually shifting ability distribution—simulating the case where candidates over time are improving/have higher ability scores; Tables 5.17 to 5.24 and Figures 5.12 to 5.15 contain the statistical results for the abrupt shift in ability distributions. This might represent the situation where all of a sudden many items were exposed to candidates at a website. All of the findings reported above for the normal distribution were observed again. Basically, the two IRT based statistics performed very well in all situations and they were very similar.

The range of raw residuals is obviously from -1 to +1 but the standardized residuals can be as big as more than 12 when item exposure exists (see Figure 5.8). The standardized residuals showed some advantages over raw residuals.

Looking at the big picture, and by-passing some of the irregularities and minor trends in the findings, we were struck by the similarity of results for the two IRT-based exposure detection statistics across the three ability distributions compared to the very different results observed with the moving average p-value statistic.

5.4 Conclusions

The results from the study were revealing for all of the variables studied: (1) ability shifts, (2) item exposure models, (3) item exposure detection statistics, and (4) item statistics. First, the ability shifts were consequential. As a starter, it was easy to see that the moving p values produced unacceptable results when shifts in the ability distribution took place over the testing window. The plots were deleted from the current version but it is easy to image the situation —basically **all** items would be flagged with shifts in the ability distribution, regardless of whether or not they were exposed. In those situations, clearly, the other two statistics would be preferred. With a normal distribution of ability over the testing window all three statistics produced comparable results.

With respect to the item exposure models, putting aside the somewhat unrealistic first case (ρ=1, 100%) where detection was easy, one finding was that the ρ=.25, 10% case produced quite unacceptable levels of exposed item detection. This is the case where 10% of the candidates have a small boost in their performance level because of prior knowledge. For an examinee with a 50% probability of success on an item, that success was upped to 62.5% under the item exposure model. For a better candidate with a probability of success of 75%, that success would be upped to 81.2%. For examinees operating at chance level based on their ability (25%) that probability would be increased to 43.75%, far from any assurance of a

correct response to the item. And in this condition, these increased probabilities would be applied to the item level performance of only 10% of the candidates. Clearly, this level of exposure would be very difficult to spot in practice. The levels of detection of exposure were substantially higher in the other two cases, but especially so for the case ρ=.25 and 100% exposure. How realistic this case might be in practice is not certain, but the detection rates were quite good, and certainly preferable to not taking any action at all.

As for the item exposure detection statistics, the result showed a strong advantage to the two IRT-based statistics. They were applicable across all conditions simulated whereas the item p-value was not. And, they typically identified exposed items except in the cases where a small amount of exposure was simulated. It is noticeable as well that what ever the detection rates, it was always easiest to detect the more difficult items, and generally the more discriminating items. Some reversals were seen in the data however.

A potential shortcoming of simulation study 2 is that the determination of the control limits may not be proper for some situations. We found the detecting statistics performed poorly for some situations. For example, for ρ=.25 and 10% case the power for all statistics is lower than 20% but meanwhile the type I error ranged around 3%. If we narrowed the control limits it is expected that the performance of the statistics would be better.

Interestingly and importantly, the findings about the item exposure detection statistics and how they functioned are applicable to all forms of computer-based testing from linear or linear-on-to-fly to multi-stage, to fully adaptive tests. Once an item is administered in whatever design is operative in the testing program, the candidate performance data can be added to the string of data being collected on

each item, and the item detection statistics can be updated, and tested for significance. An item remains in the bank until it is retired or identified as being exposed. The likelihood of detection of exposed items obviously depends on the confidence bands that have been established (which depend on the window size, in this study the number of candidates used in the statistics was 100), the statistical characteristics of the test items, and the type of exposure taking place. For the two IRT-based statistics, that considered ability in the calculation of statistics, the nature of the ability distribution was irrelevant. Test administrators should be pleased to discover that the harder and more discriminating items are the ones that can be detected fastest. These are the same items that influence the ability estimates the most, and therefore they raise the most questions about the validity of candidate scores.

Table 5.1. Number of times of item administration after exposure. (ρ = 1.0, for 100%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Moving P values	b=-1.00	27.4	22.0	28.6
	b= 0.00	15.5	10.4	9.0
	b= 1.00	11.9	7.3	4.5
	b= 2.00	9.2	4.7	2.6
Moving Item Residuals	b=-1.00	25.3	22.9	24
	b= 0.00	16.3	12.4	11.2
	b= 1.00	12.5	8.7	7.5
	b= 2.00	10.4	6.4	3.6
Standardized Item Residuals	b=-1.00	25.2	22.6	23.5
	b= 0.00	16.3	12.4	10.9
	b= 1.00	12.4	8.6	7.5
	b= 2.00	10.4	6.7	4.6

Table 5.2. Type I errors and power. (ρ = 1.0, for 100%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving P Values	b=-1.00	1.50	100.0	3.36	100.0	2.33	100.0
	b= 0.00	2.68	100.0	4.42	100.0	3.60	100.0
	b= 1.00	2.16	100.0	2.86	100.0	5.55	100.0
	b= 2.00	1.99	100.0	4.08	100.0	6.61	100.0
Moving Item Residuals	b=-1.00	2.14	100.0	2.78	100.0	1.97	100.0
	b= 0.00	2.55	100.0	3.27	100.0	1.94	100.0
	b= 1.00	2.36	100.0	2.56	100.0	2.85	100.0
	b= 2.00	2.02	100.0	2.63	100.0	3.11	100.0
Standardized Item Residuals	b=-1.00	2.15	100.0	2.78	100.0	2.16	100.0
	b= 0.00	2.55	100.0	3.10	100.0	1.94	100.0
	b= 1.00	2.45	100.0	2.74	100.0	2.88	100.0
	b= 2.00	2.09	100.0	2.59	100.0	2.99	100.0

Table 5.3. Number of times of item administration after exposure. ($\rho = 1.0$, for 10%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Moving P values	b=-1.00	320.7	301.2	292.3
	b= 0.00	173.8	160.2	140.3
	b= 1.00	169.0	115.9	61.5
	b= 2.00	98.5	57.2	44.8
Moving Item Residuals	b=-1.00	283.7	313.9	329.8
	b= 0.00	191.7	143.5	188.5
	b= 1.00	140.8	113.8	66.6
	b= 2.00	98.1	61.2	48.8
Standardized Item Residuals	b=-1.00	283.8	315.4	307.2
	b= 0.00	192.9	149.1	189.5
	b= 1.00	135.0	112.4	67.8
	b= 2.00	96.9	60.8	48.8

Table 5.4. Type I errors and power. ($\rho = 1.0$, for 10%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving P Values	b=-1.00	1.50	8.3	3.36	10.5	2.33	9.80
	b= 0.00	2.68	16.8	4.41	23.7	3.63	25.2
	b= 1.00	2.16	26.6	2.86	38.4	5.55	64.7
	b= 2.00	1.99	47.5.	4.08	77.9	6.60	94.7
Moving Item Residuals	b=-1.00	2.14	9.8	2.78	10.0	1.97	8.7
	b= 0.00	2.55	16.7	3.27	23.8	1.94	24.1
	b= 1.00	2.36	29.5	2.56	41.1	2.84	63.6
	b= 2.00	2.02	49.0	2.62	75.5	3.10	94.0
Standardized Item Residuals	b=-1.00	2.15	9.9	2.78	10.0	2.16	9.1
	b= 0.00	2.54	16.7	3.10	23.4	1.94	24.3
	b= 1.00	2.45	29.9	2.74	42.2	2.88	63.5
	b= 2.00	2.08	49.4	2.59	74.9	2.99	93.5

Table 5.5. Number of times of item administration after exposure. ($\rho = 0.25$, for 100%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Moving P values	b=-1.00	113.5	123.5	118.8
	b= 0.00	67.9	55.3	55.4
	b= 1.00	53.8	49.7	24.5
	b= 2.00	39.5	21.0	16.1
Moving Item Residuals	b=-1.00	99.4	119.5	109.9
	b= 0.00	64.9	52.6	56.0
	b= 1.00	47.1	46.2	29.6
	b= 2.00	38.4	23.3	18.7
Standardized Item Residuals	b=-1.00	99.3	119.1	109.3
	b= 0.00	64.9	52.9	56.0
	b= 1.00	46.3	45.6	29.7
	b= 2.00	38.3	25.1	20.8

Table 5.6. Type I errors and power. ($\rho = 0.25$, for 100%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving P Values	b=-1.00	1.50	40.9	3.36	39.0	2.33	33.9
	b= 0.00	2.68	71.6	4.41	78.0	3.63	85.5
	b= 1.00	2.16	88.8	2.86	97.3	5.55	99.8
	b= 2.00	1.99	99.3	4.08	100.0	6.60	100.0
Moving Item Residuals	b=-1.00	2.14	46.7	2.78	39.5	1.97	41.0
	b= 0.00	2.55	74.0	3.27	80.8	1.94	89.1
	b= 1.00	2.36	91.2	2.56	98.2	2.84	99.9
	b= 2.00	2.02	99.4	2.62	100.0	3.10	100.0
Standardized Item Residuals	b=-1.00	2.15	47.2	2.78	39.8	2.16	42.0
	b= 0.00	2.54	74.0	3.10	80.5	1.94	89.2
	b= 1.00	2.45	91.4	2.74	98.3	2.88	99.8
	b= 2.00	2.08	99.4	2.59	100.0	2.99	100.0

Table 5.7. Number of times of item administration after exposure. ($\rho = 0.25$, for 10%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Moving P values	b=-1.00	517.6	473.3	393.2
	b= 0.00	530.9	420.6	310.1
	b= 1.00	539.2	340.4	186.2
	b= 2.00	424.2	173.1	136.8
Moving Item Residuals	b=-1.00	666.5	622.5	721.6
	b= 0.00	482.3	538.2	478.2
	b= 1.00	558.9	415.1	270.0
	b= 2.00	480.9	271.6	179.3
Standardized Item Residuals	b=-1.00	650.5	671.9	674.7
	b= 0.00	482.9	591.6	479.0
	b= 1.00	573.2	388.0	282.3
	b= 2.00	474.4	255.3	180.5

Table 5.8. Type I errors and power. ($\rho = 0.25$, for 10%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving P Values	b=-1.00	1.50	3.34	3.36	4.2	2.33	4.5
	b= 0.00	2.68	4.53	4.41	7.6	3.63	8.0
	b= 1.00	2.16	5.34	2.86	7.7	5.55	16.0
	b= 2.00	1.99	6.81	4.08	15.5	6.60	26.1
Moving Item Residuals	b=-1.00	2.14	3.72	2.78	3.4	1.97	3.5
	b= 0.00	2.55	4.68	3.27	6.0	1.94	4.8
	b= 1.00	2.36	6.12	2.56	7.4	2.84	10.8
	b= 2.00	2.02	7.03	2.62	11.8	3.10	21.1
Standardized Item Residuals	b=-1.00	2.15	3.78	2.78	3.4	2.16	3.6
	b= 0.00	2.54	4.67	3.10	5.8	1.94	4.9
	b= 1.00	2.45	6.23	2.74	7.8	2.88	10.6
	b= 2.00	2.08	7.24	2.59	11.5	2.99	20.6

Table 5.9. Number of times of item administration after exposure. (ρ = 1.0, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	23.2	22.7	26.3
	b= 0.00	16.4	13.5	11.9
	b= 1.00	12.8	10.4	9.0
	b= 2.00	10.7	6.8	4.1
Standardized Item Residuals	b=-1.00	23.2	22.8	26.7
	b= 0.00	16.8	13.4	12.1
	b= 1.00	13.1	10.4	9.0
	b= 2.00	10.9	7.8	4.3

Table 5.10. Type I errors and power. (ρ = 1.0, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	3.4	100.0	1.9	100.0	3.4	100.0
	b= 0.00	3.5	100.0	2.9	100.0	3.4	100.0
	b= 1.00	2.8	100.0	1.9	100.0	2.4	100.0
	b= 2.00	2.2	100.0	1.4	100.0	2.2	100.0
Standardized Item Residuals	b=-1.00	3.1	100.0	1.5	100.0	2.4	100.0
	b= 0.00	3.2	100.0	2.8	100.0	2.7	100.0
	b= 1.00	2.5	100.0	2.2	100.0	2.5	100.0
	b= 2.00	2.3	100.0	2.0	100.0	3.6	100.0

Table 5.11. Number of times of item administration after exposure. (ρ = 1.0, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	285.3	312.3	278.2
	b= 0.00	182.7	163.6	133.1
	b= 1.00	130.1	88.5	73.5
	b= 2.00	106.2	74.9	47.3
Standardized Item Residuals	b=-1.00	280.0	304.4	320.9
	b= 0.00	190.0	170.5	149.1
	b= 1.00	134.8	89.4	74.1
	b= 2.00	114.2	73.7	47.9

Table 5.12. Type I errors and power. (ρ = 1.0, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	3.4	5.1	1.9	3.8	3.4	2.4
	b= 0.00	3.5	12.7	2.9	14.2	3.4	12.8
	b= 1.00	2.8	24.2	1.9	30.6	2.4	48.7
	b= 2.00	2.2	45.9	1.4	66.0	2.2	86.9
Standardized Item Residuals	b=-1.00	3.1	7.1	1.5	6.9	2.4	6.1
	b= 0.00	3.2	12.1	2.8	15.2	2.7	12.9
	b= 1.00	2.5	19.9	2.2	25.3	2.5	37.2
	b= 2.00	2.3	38.0	2.0	54.4	3.6	74.7

Table 5.13. Number of times of item administration after exposure. ($\rho = 0.25$, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	89.6	122.6	128.9
	b= 0.00	60.1	55.6	50
	b= 1.00	45.0	42.2	29.6
	b= 2.00	45.0	27.9	17.3
Standardized Item Residuals	b=-1.00	90.1	124.3	128.8
	b= 0.00	62.2	56.0	54.2
	b= 1.00	46.8	42.3	30.2
	b= 2.00	45.6	28.4	17.7

Table 5.14. Type I errors and power. ($\rho = 0.25$, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	3.4	28.7	1.9	12.2	3.4	7.30
	b= 0.00	3.5	58.8	2.9	55.7	3.4	56.94
	b= 1.00	2.8	85.0	1.9	92.0	2.4	98.03
	b= 2.00	2.2	98.0	1.4	99.9	2.2	100.0
Standardized Item Residuals	b=-1.00	3.1	33.7	1.5	19.5	2.4	16.3
	b= 0.00	3.2	58.0	2.8	57.7	2.7	57.4
	b= 1.00	2.5	81.3	2.2	89.4	2.5	95.7
	b= 2.00	2.3	96.8	2.0	99.5	3.6	100.0

Table 5.15. Number of times of item administration after exposure. ($\rho = 0.25$, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	586.4	499.0	397.9
	b= 0.00	470.7	496.0	348.8
	b= 1.00	449.8	462.3	282.3
	b= 2.00	364.8	282.9	169.4
Standardized Item Residuals	b=-1.00	609.8	528.9	518.7
	b= 0.00	548.5	497.7	375.9
	b= 1.00	477.2	493.0	296.5
	b= 2.00	409.6	306.9	177.6

Table5.16. Type I errors and power. ($\rho = 0.25$, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	3.4	1.9	1.9	1.3	3.4	1.1
	b= 0.00	3.5	3.0	2.9	4.5	3.4	4.3
	b= 1.00	2.8	5.3	1.9	7.9	2.4	11.3
	b= 2.00	2.2	9.7	1.4	12.1	2.2	23.6
Standardized Item Residuals	b=-1.00	3.1	2.7	1.5	2.6	2.4	3.1
	b= 0.00	3.2	2.8	2.8	4.9	2.7	4.4
	b= 1.00	2.5	4.1	2.2	5.8	2.5	6.9
	b= 2.00	2.3	6.4	2.0	6.5	3.6	12.3

Table 5.17. Number of times of item administration after exposure. (ρ = 1.0, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	38.9	48.9	63.9
	b= 0.00	23.6	22.3	24.9
	b= 1.00	15.4	14.3	10.1
	b= 2.00	12.7	7.7	4.9
Standardized Item Residuals	b=-1.00	40.6	50.5	70.0
	b= 0.00	23.5	22.3	25.3
	b= 1.00	15.0	12.5	8.0
	b= 2.00	11.2	5.3	2.2

Table 5.18. Type I errors and power. (ρ = 1.0, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	4.1	100.0	4.0	100.0	6.3	77.8
	b= 0.00	2.2	100.0	3.9	100.0	3.2	100.0
	b= 1.00	1.9	100.0	1.3	100.0	0.5	100.0
	b= 2.00	0.9	100.0	0.2	100.0	0.2	100.0
Standardized Item Residuals	b=-1.00	2.7	100.0	1.7	100.0	2.2	100.0
	b= 0.00	2.3	100.0	3.9	100.0	3.3	100.0
	b= 1.00	2.6	100.0	3.3	100.0	2.9	100.0
	b= 2.00	2.0	100.0	3.5	100.0	4.2	100.0

Table 5.19. Number of times of item administration after exposure. (ρ = 1.0, for 10%, abrupt change in the mean of the ability distribution)

		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	495.2	532.0	355
	b= 0.00	249.3	206.4	271.9
	b= 1.00	162.7	148.1	95.5
	b= 2.00	131.4	80.1	54.5
Standardized Item Residuals	b=-1.00	413.9	601.8	669.4
	b= 0.00	248.9	236.2	274.2
	b= 1.00	206.5	194.3	116.0
	b= 2.00	149.9	95.0	57.4

Table 5.20. Type I errors and power. (ρ = 1.0, for 10%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	4.1	1.6	4.0	1.5	6.3	0.2
	b= 0.00	2.2	8.8	3.9	9.9	3.2	7.1
	b= 1.00	1.9	24.1	1.3	28.0	0.5	42.3
	b= 2.00	0.9	52.0	0.2	61.1	0.2	82.2
Standardized Item Residuals	b=-1.00	2.7	6.4	1.7	4.5	2.2	3.4
	b= 0.00	2.3	9.4	3.9	9.9	3.3	7.6
	b= 1.00	2.6	14.7	3.3	19.7	2.9	26.5
	b= 2.00	2.0	32.5	3.5	43.7	4.2	60.6

Table 5.21. Number of times of item administration after exposure. ($\rho = 0.25$, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	183.6	333.8	652.1
	b= 0.00	105.8	88.7	102
	b= 1.00	55.7	60.5	41.0
	b= 2.00	50.4	32.8	24.0
Standardized Item Residuals	b=-1.00	173.4	234.0	332.5
	b= 0.00	104.7	89.3	103.3
	b= 1.00	57.5	64.9	38.8
	b= 2.00	50.9	26.6	17.7

Table 5.22. Type I errors and power. ($\rho = 0.25$, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	4.1	19.9	4.0	6.7	6.3	1.4
	b= 0.00	2.2	44.5	3.9	41.3	3.2	35.7
	b= 1.00	1.9	76.2	1.3	86.0	0.5	94.3
	b= 2.00	0.9	96.7	0.2	99.4	0.2	100.0
Standardized Item Residuals	b=-1.00	2.7	25.6	1.7	15.6	2.2	8.7
	b= 0.00	2.3	45.8	3.9	41.7	3.3	37.1
	b= 1.00	2.6	69.5	3.3	79.7	2.9	86.8
	b= 2.00	2.0	93.9	3.5	97.8	4.2	100.0

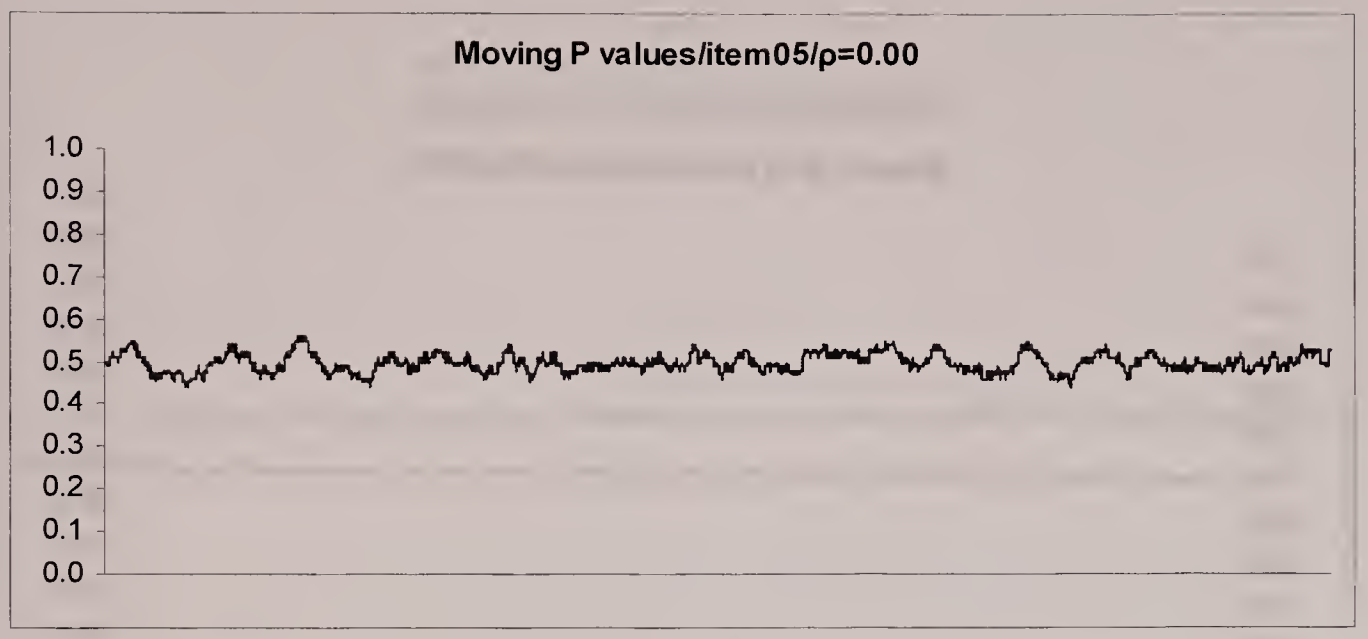
Table 5.23. Number of times of item administration after exposure. ($\rho = 0.25$, for 10%, abrupt change in the mean of the ability distribution)

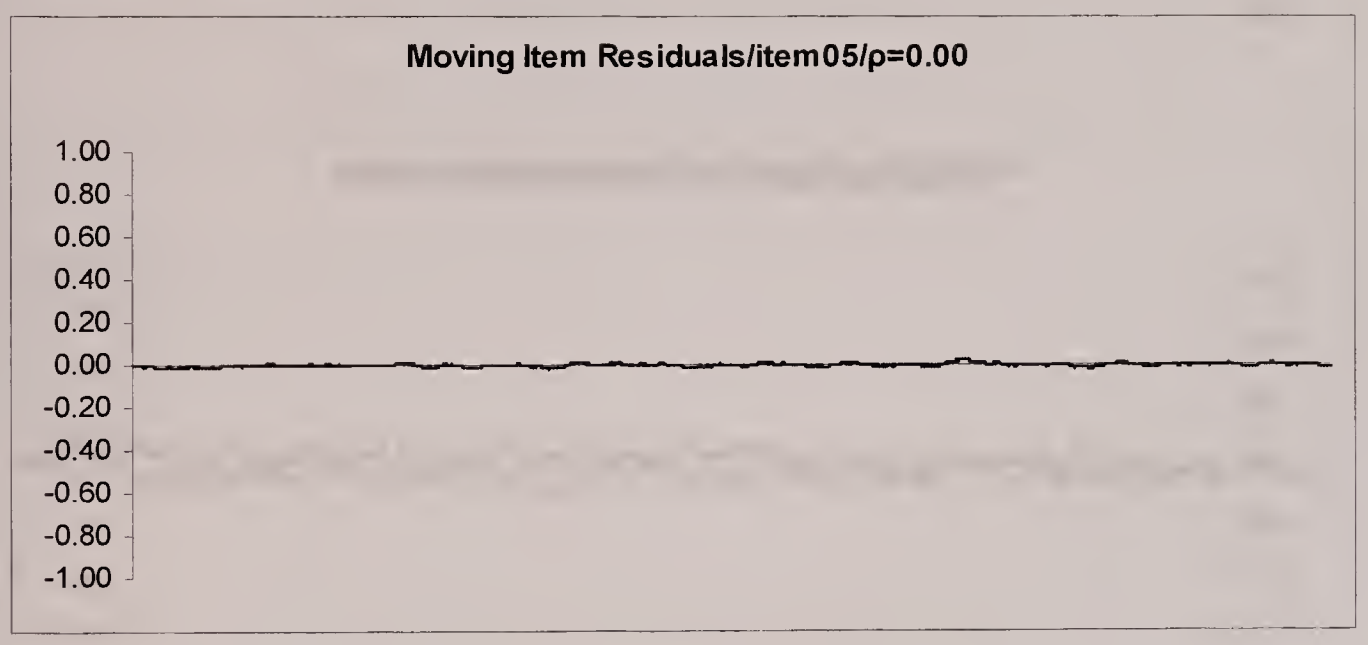
		a=0.40	a=0.70	a=1.20
Moving Item Residuals	b=-1.00	870.9	538.6	222.6
	b= 0.00	579.7	468.1	539.1
	b= 1.00	438.0	357.0	288.4
	b= 2.00	392.7	240.6	163.7
Standardized Item Residuals	b=-1.00	768.7	719.0	659.4
	b= 0.00	525.1	466.9	556.5
	b= 1.00	505.8	521.8	338.9
	b= 2.00	539.9	407.7	328.7

Table 5.24. Type I errors and power. ($\rho = 0.25$, for 10%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Moving Item Residuals	b=-1.00	4.1	2.1	4.0	0.5	6.3	0.12
	b= 0.00	2.2	3.0	3.9	3.1	3.2	2.4
	b= 1.00	1.9	5.2	1.3	8.2	0.5	1.2
	b= 2.00	0.9	10.9	0.2	15.9	0.2	26.9
Standardized Item Residuals	b=-1.00	2.7	3.2	1.7	2.2	2.2	2.0
	b= 0.00	2.3	3.3	3.9	3.1	3.3	2.7
	b= 1.00	2.6	3.5	3.3	5.2	2.9	5.4
	b= 2.00	2.0	6.4	3.5	7.2	4.2	8.3

Figure 5.1. plot of item exposure statistics for item 5. (normal ability distribution, ρ = 0.0)





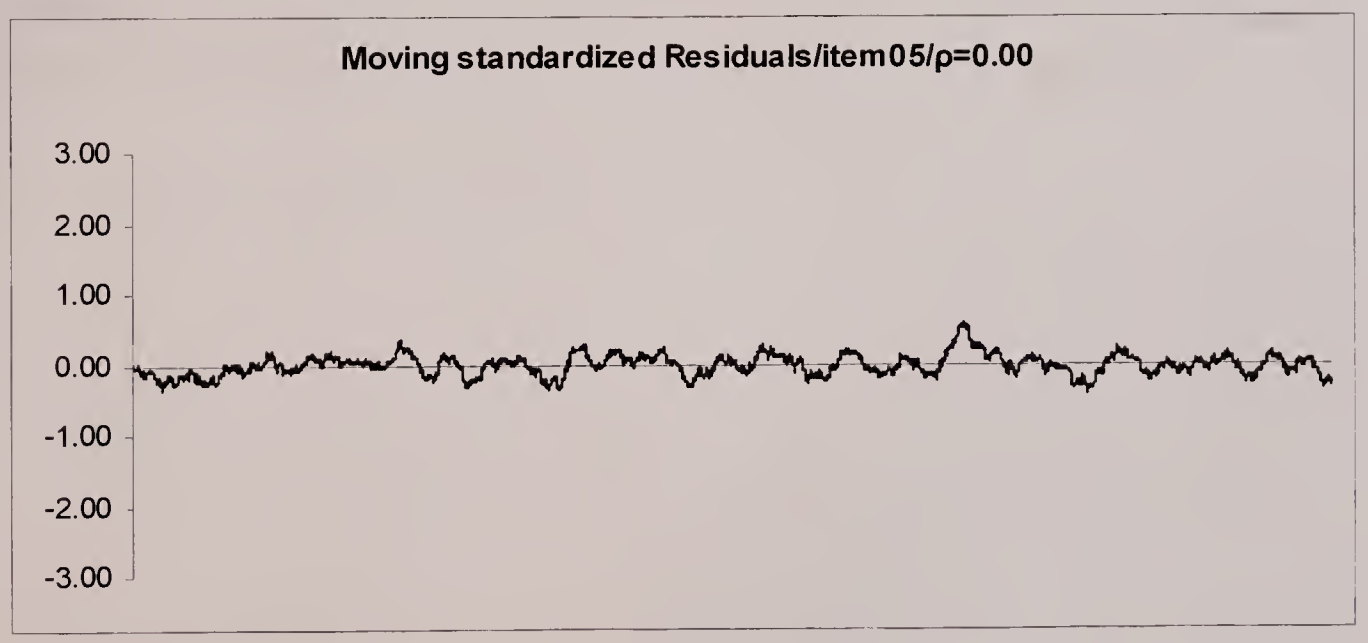
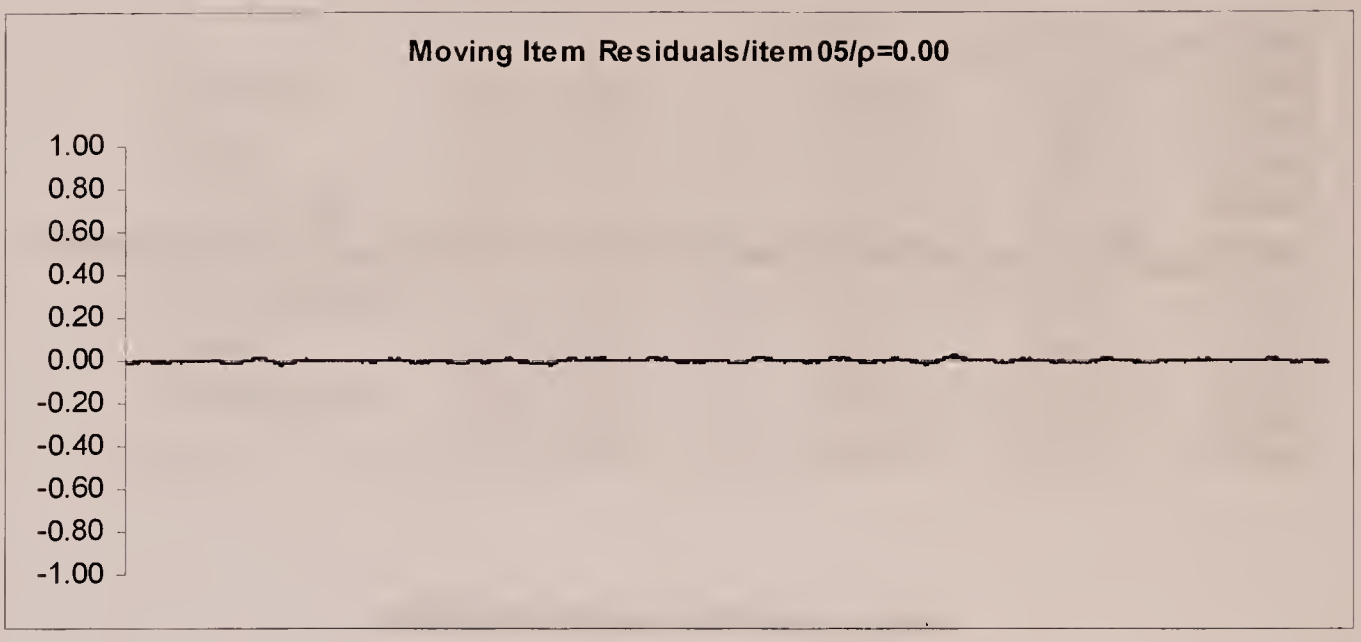


Figure 5.2. plot of item exposure statistics for item 5. (gradually shifting ability distribution, $\rho = 0.0$)



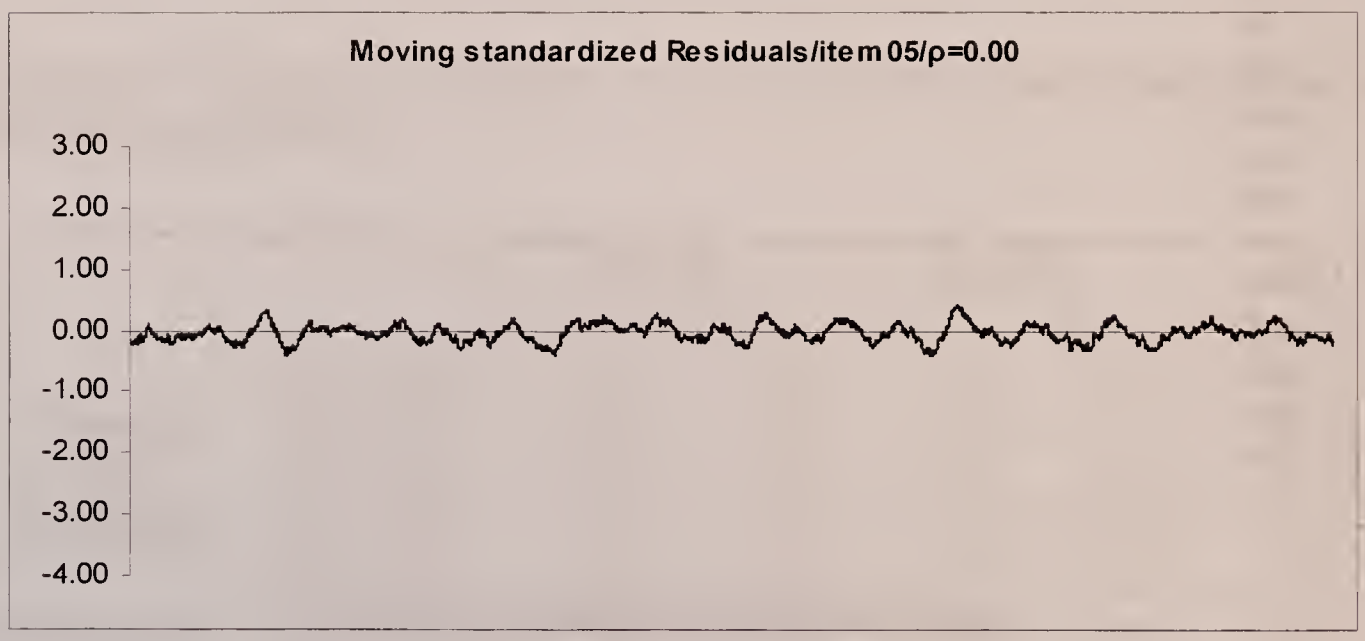
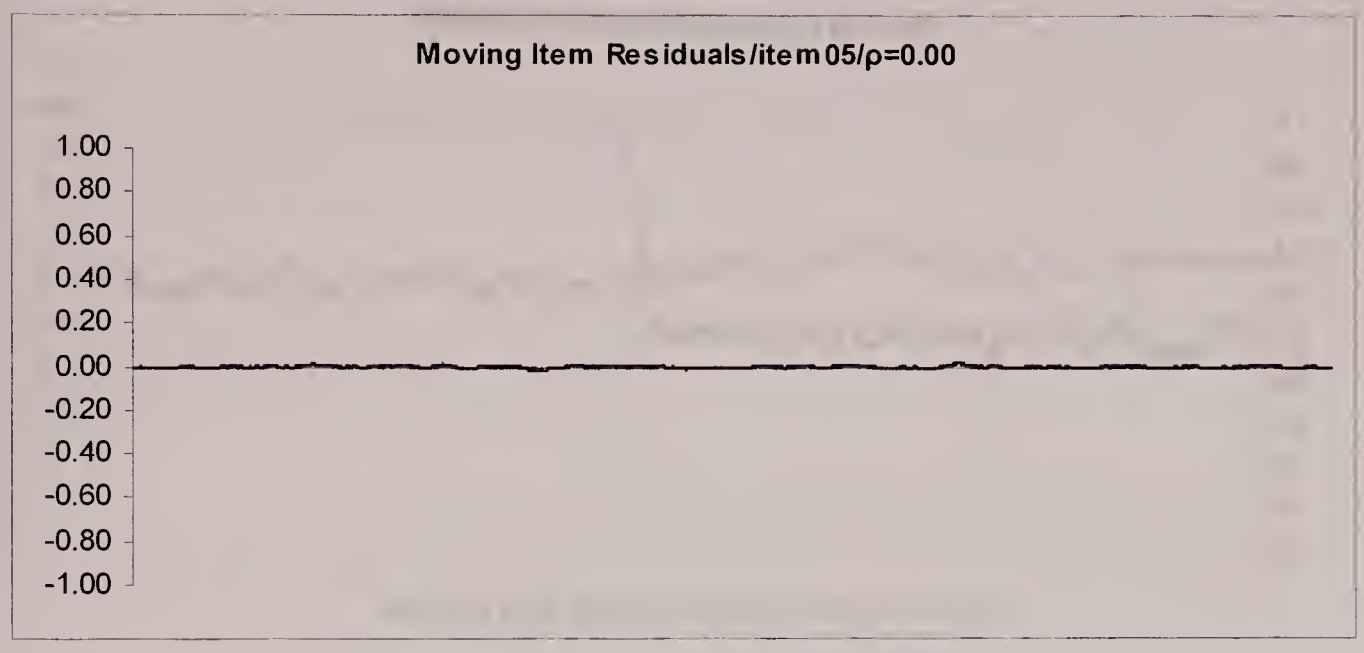


Figure 5.3. plot of item exposure statistics for item 5. (abrupt shift in ability distribution, $\rho = 0.0$)



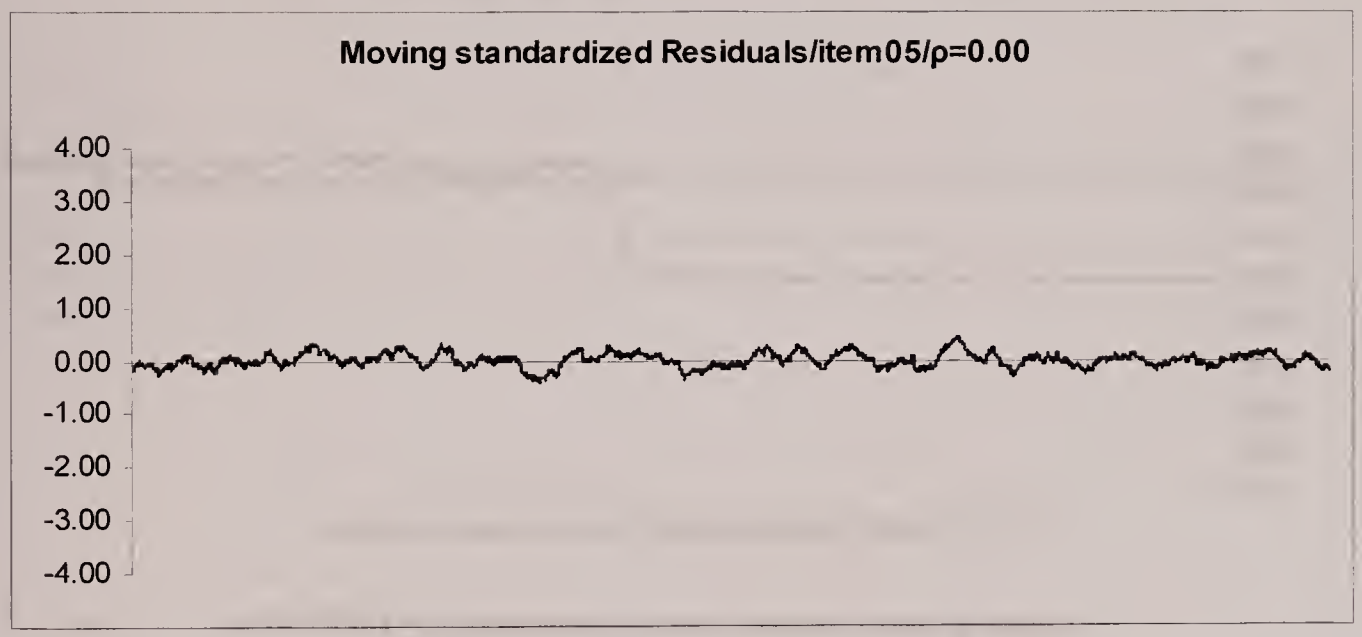
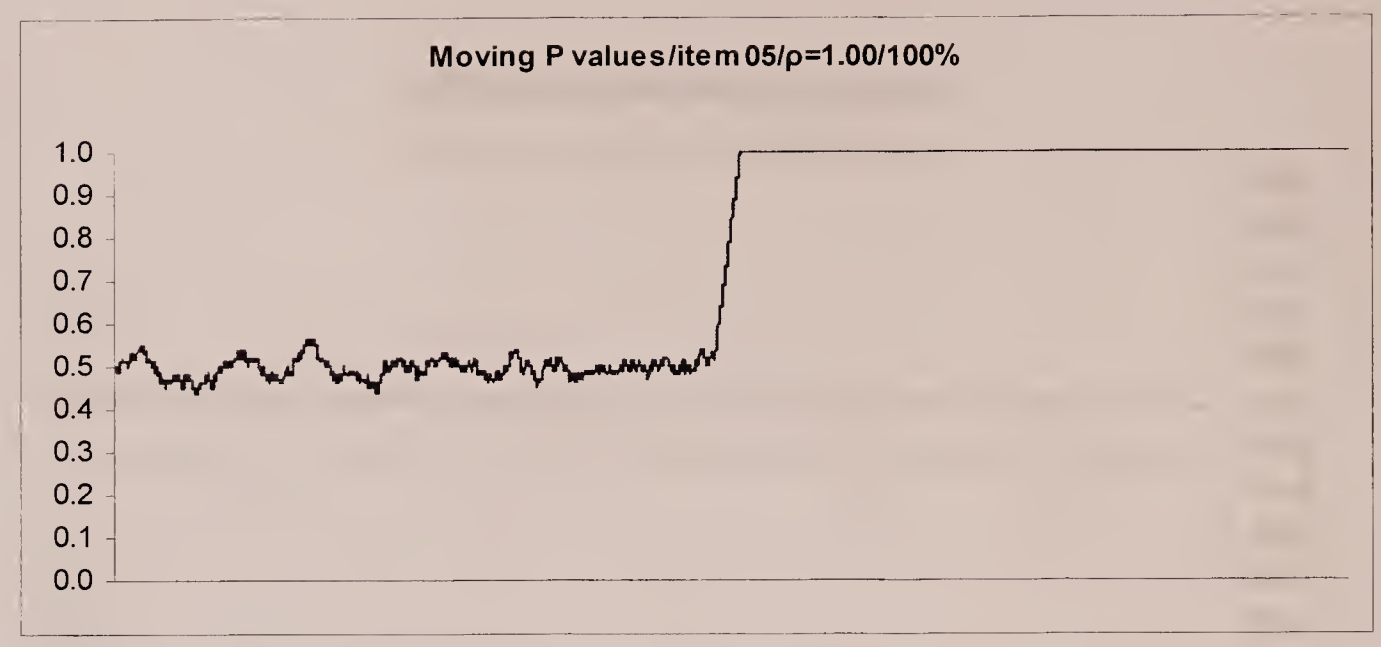
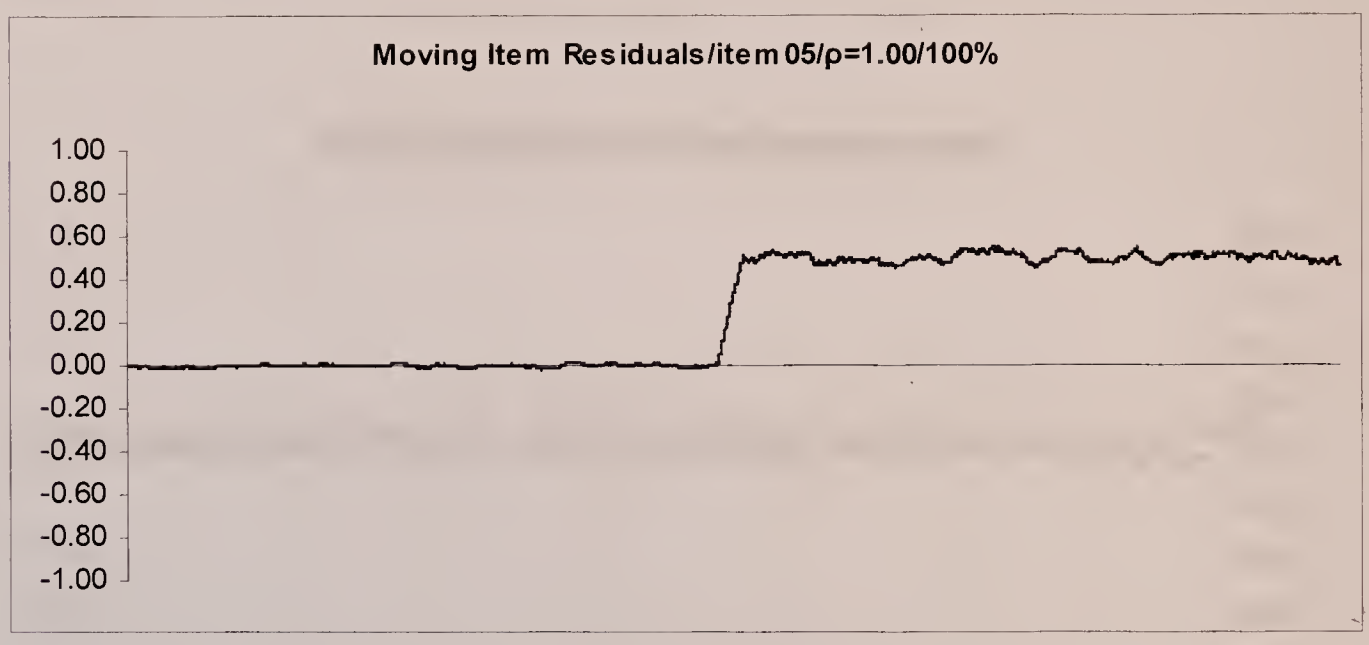


Figure 5.4. plot of item exposure statistics for item 5. (normal ability distribution, ρ = 1.0, 100%)





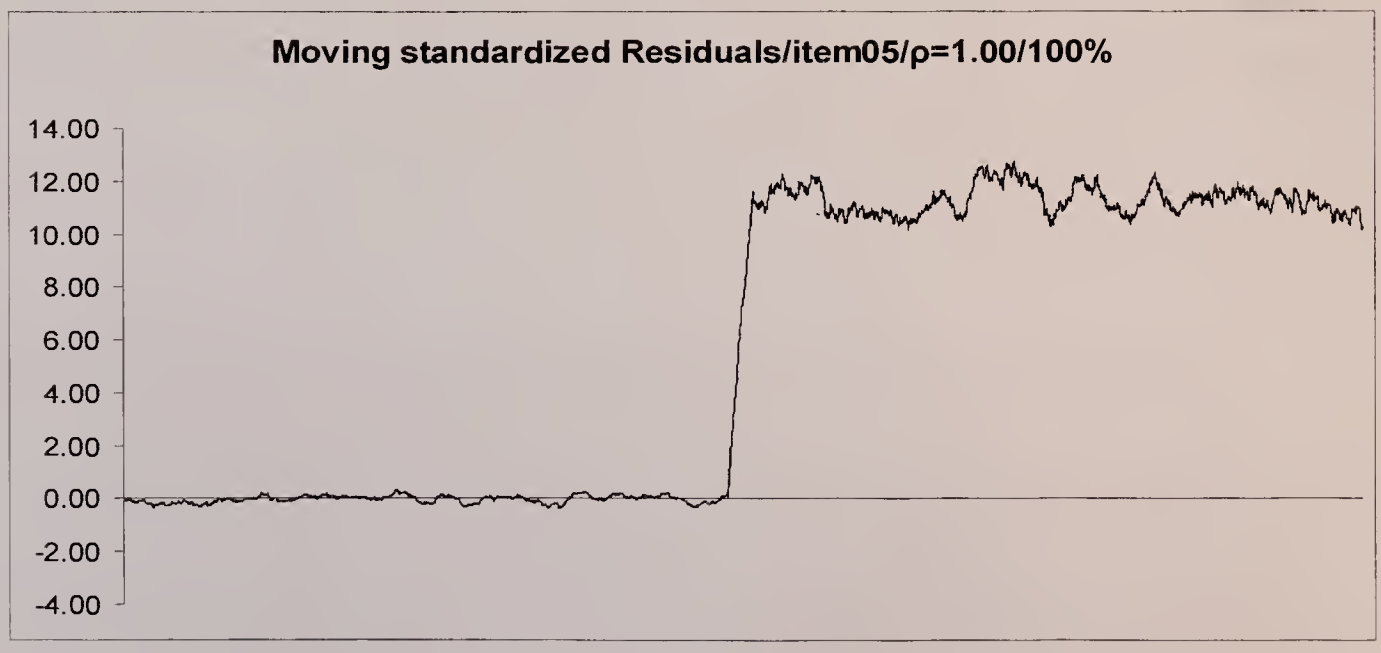
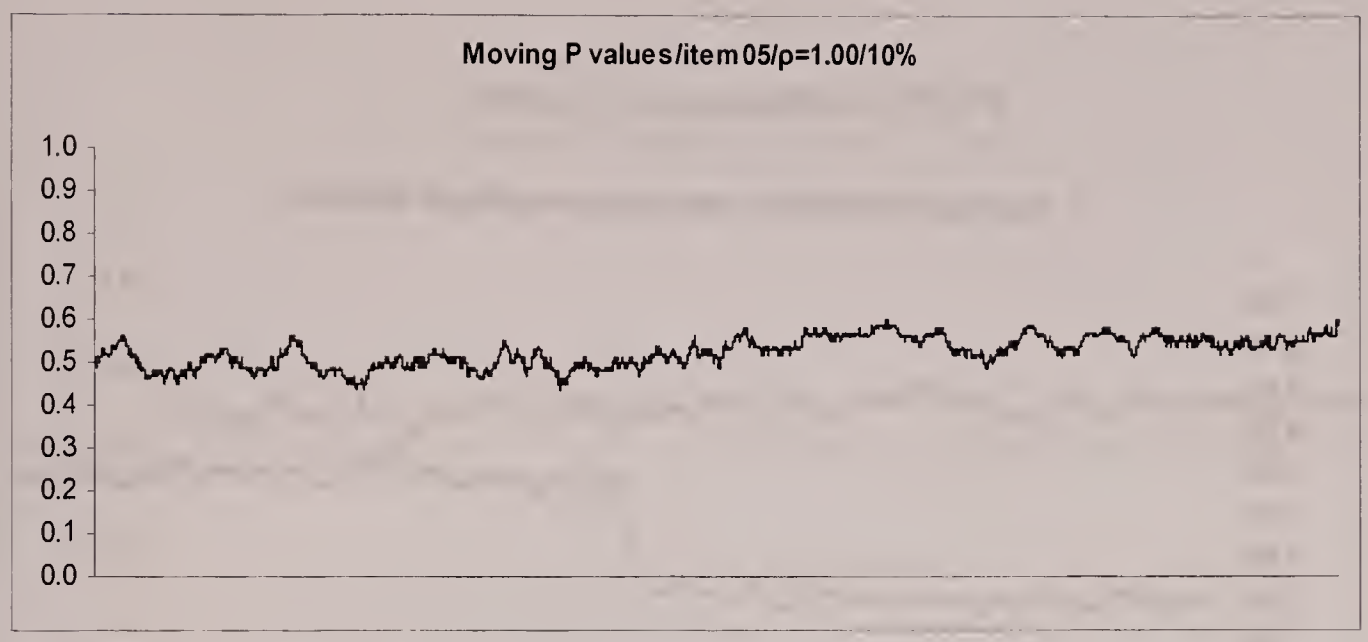
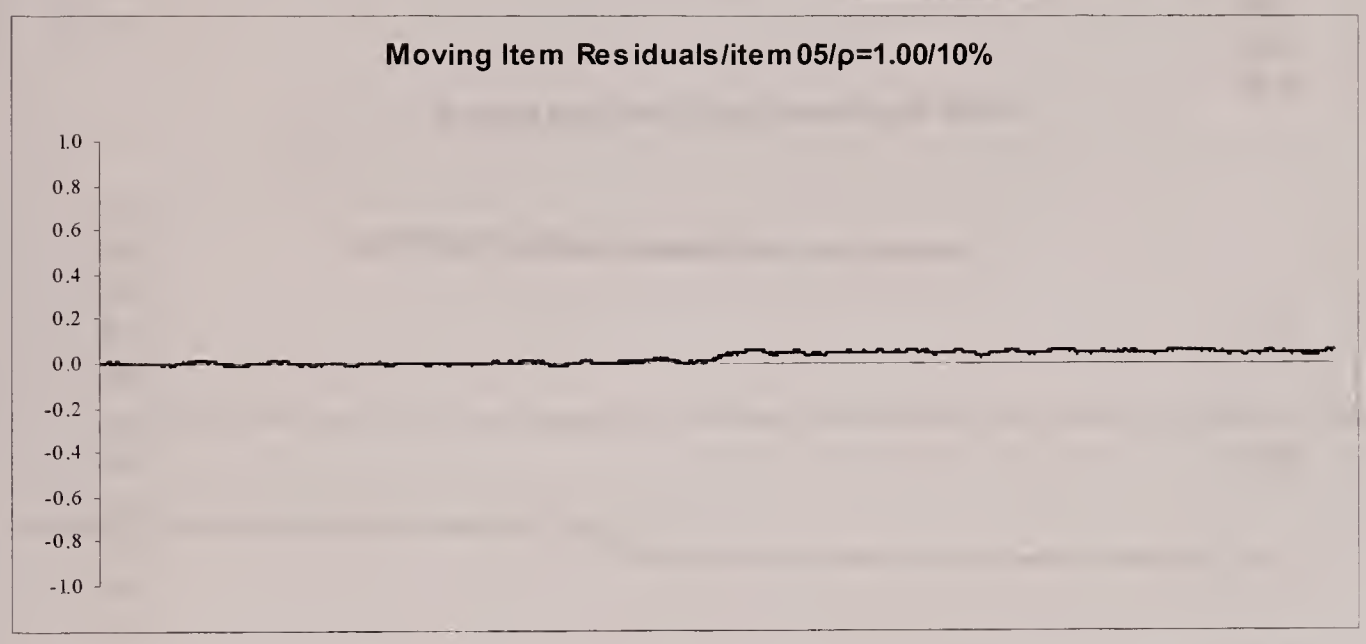


Figure 5.5. plot of item exposure statistics for item 5. (normal ability distribution, ρ = 1.0, 10%)





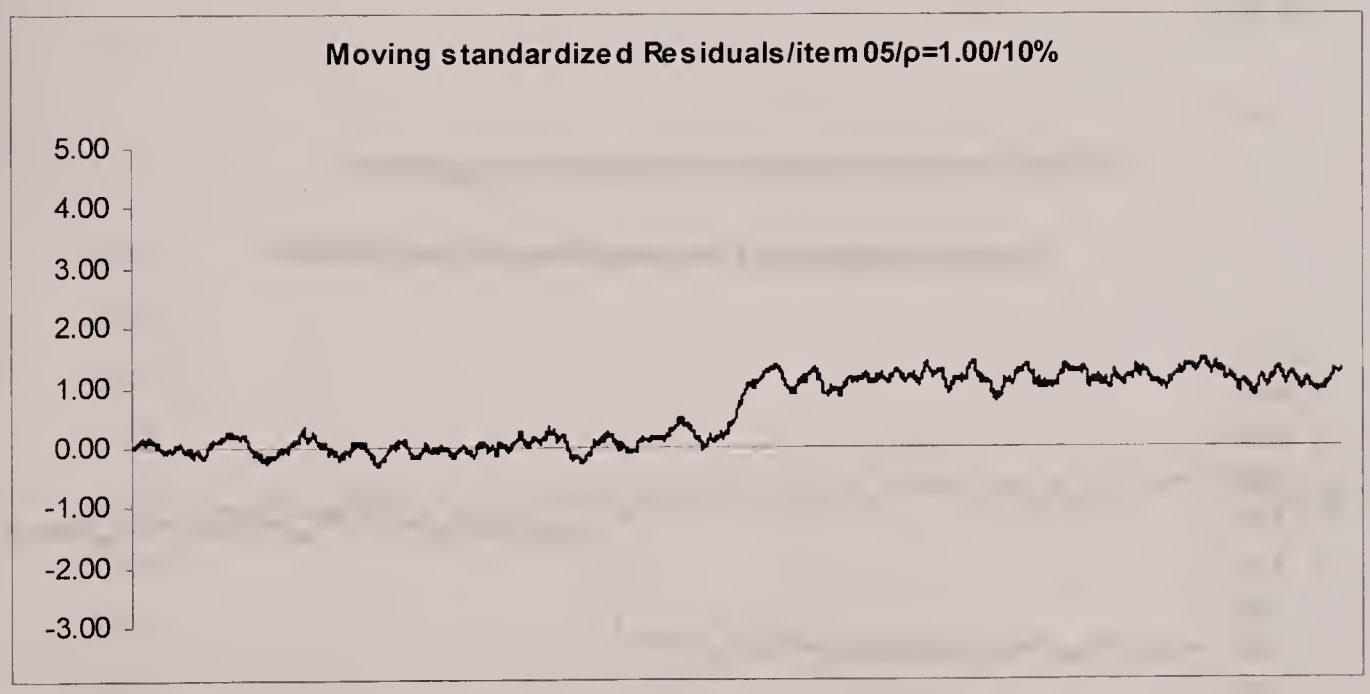
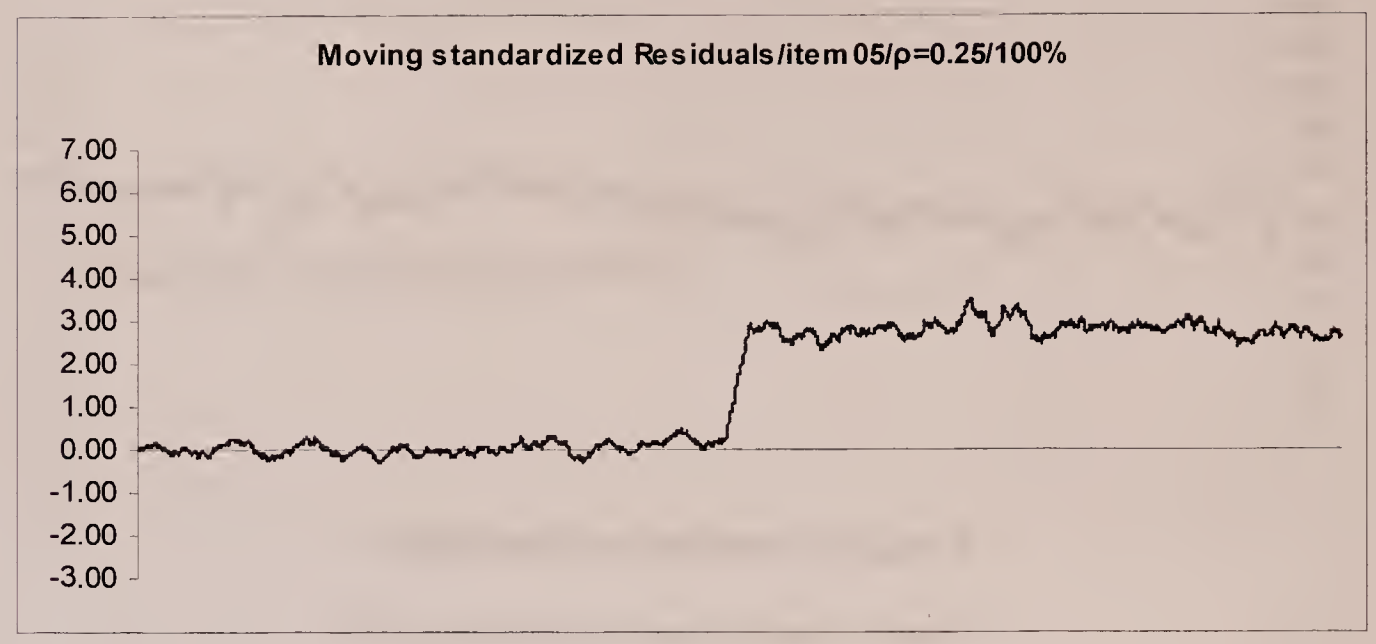
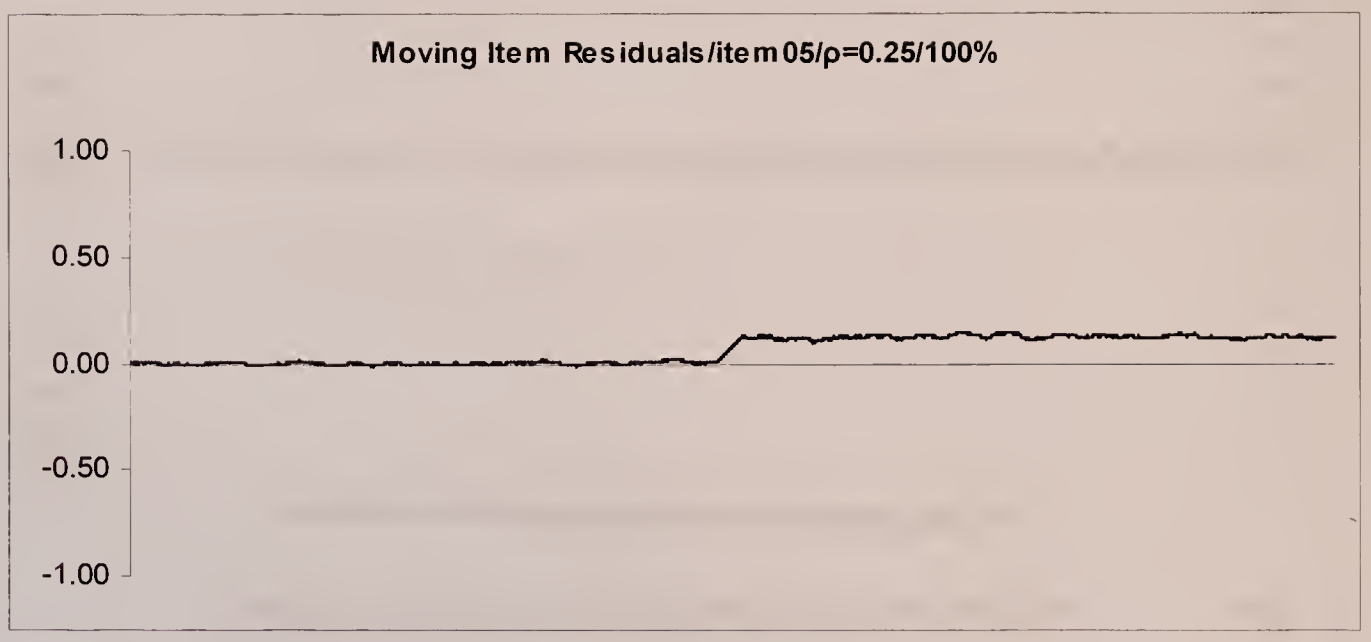


Figure 5.6. plot of item exposure statistics for item 5. (normal ability distribution, ρ = 0.25, 100%)





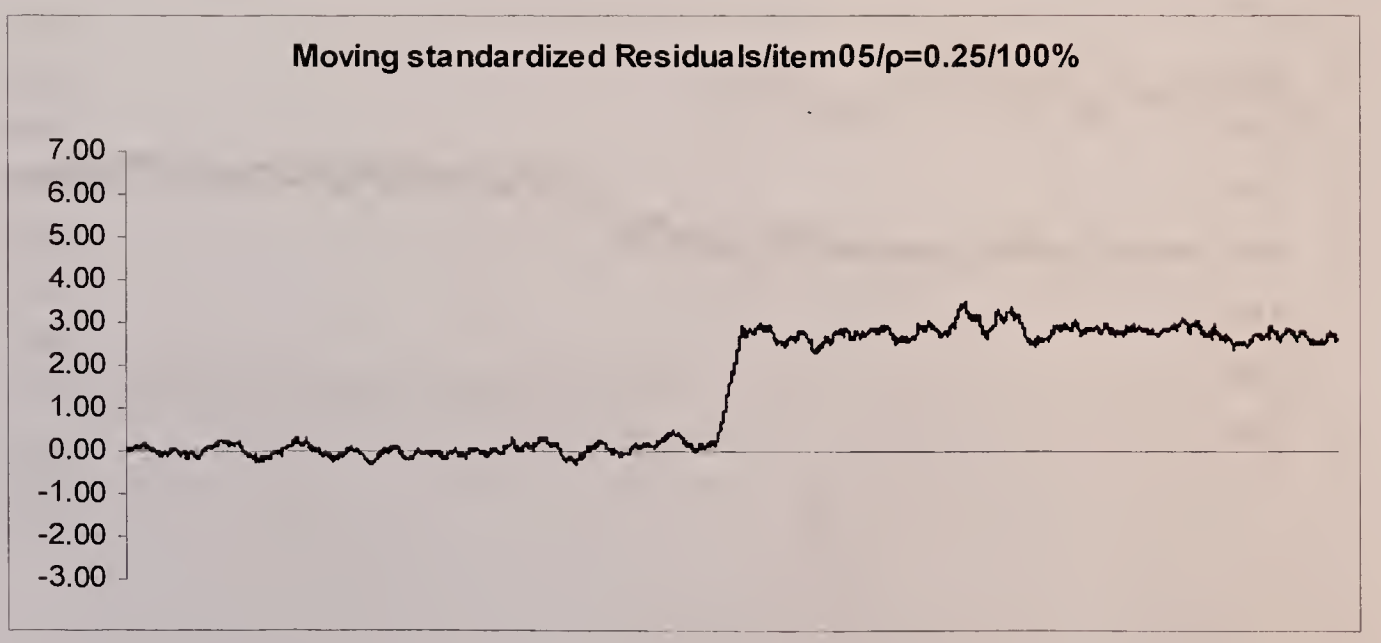
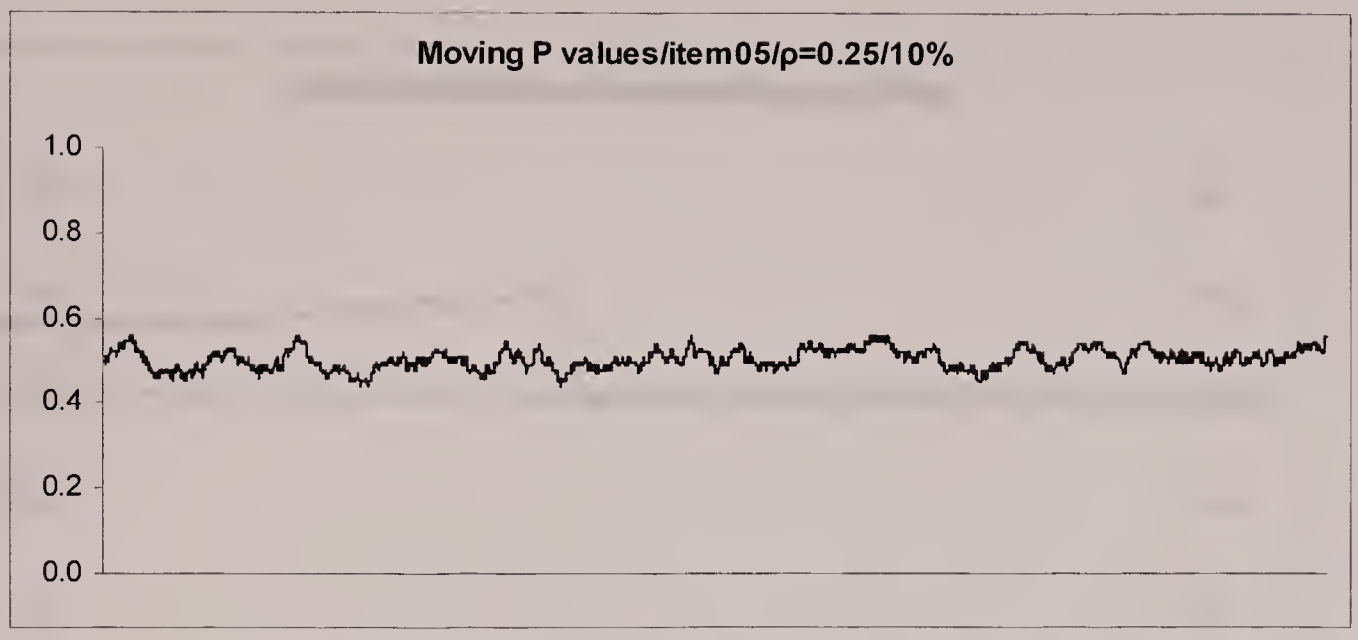
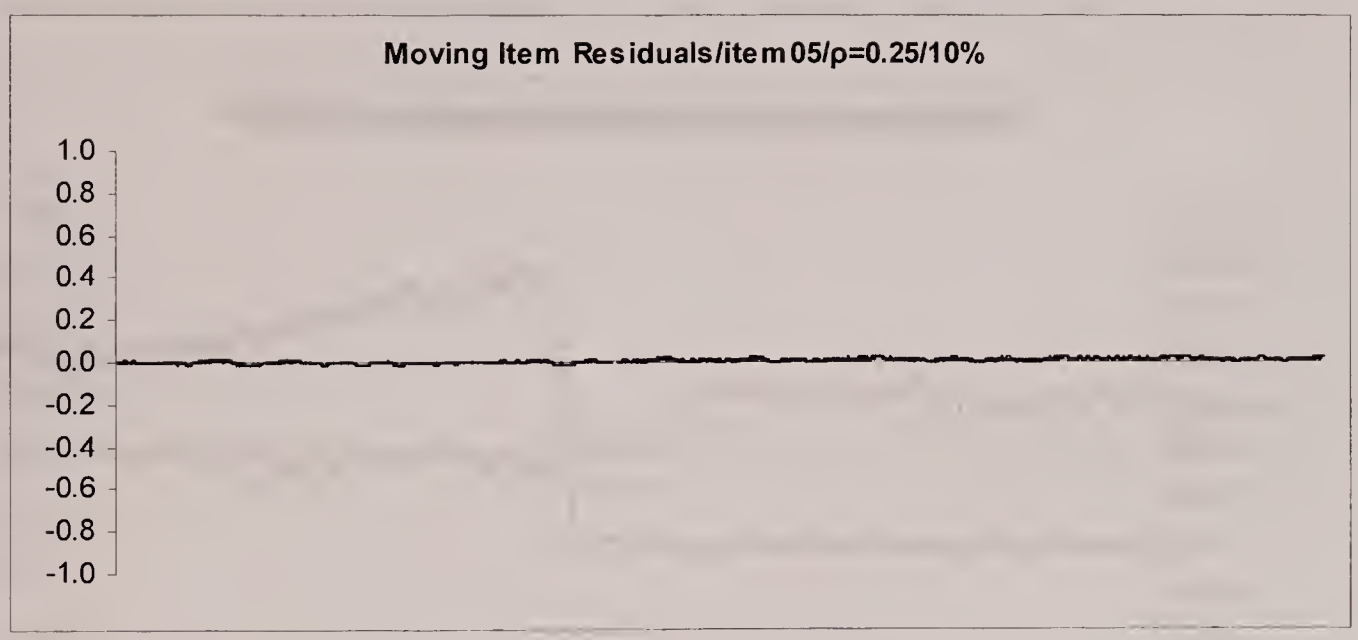


Figure 5.7. plot of item exposure statistics for item 5. (normal ability distribution, ρ = 0.25, 10%)





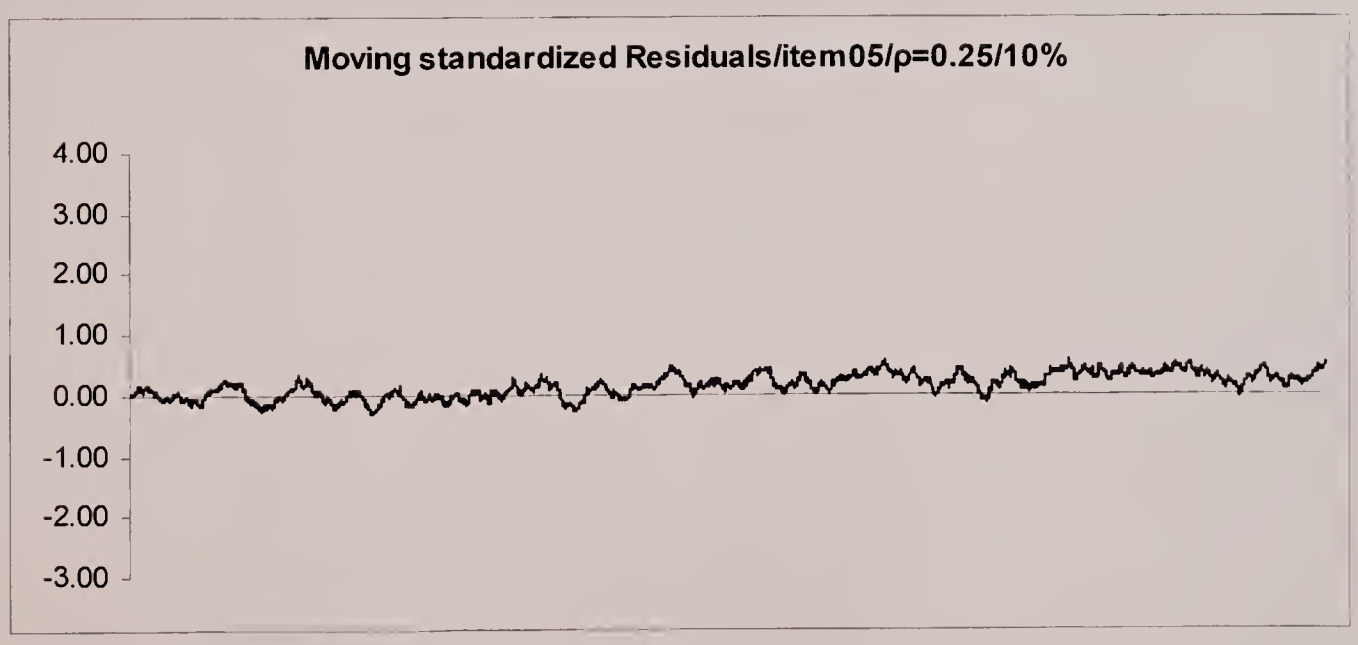
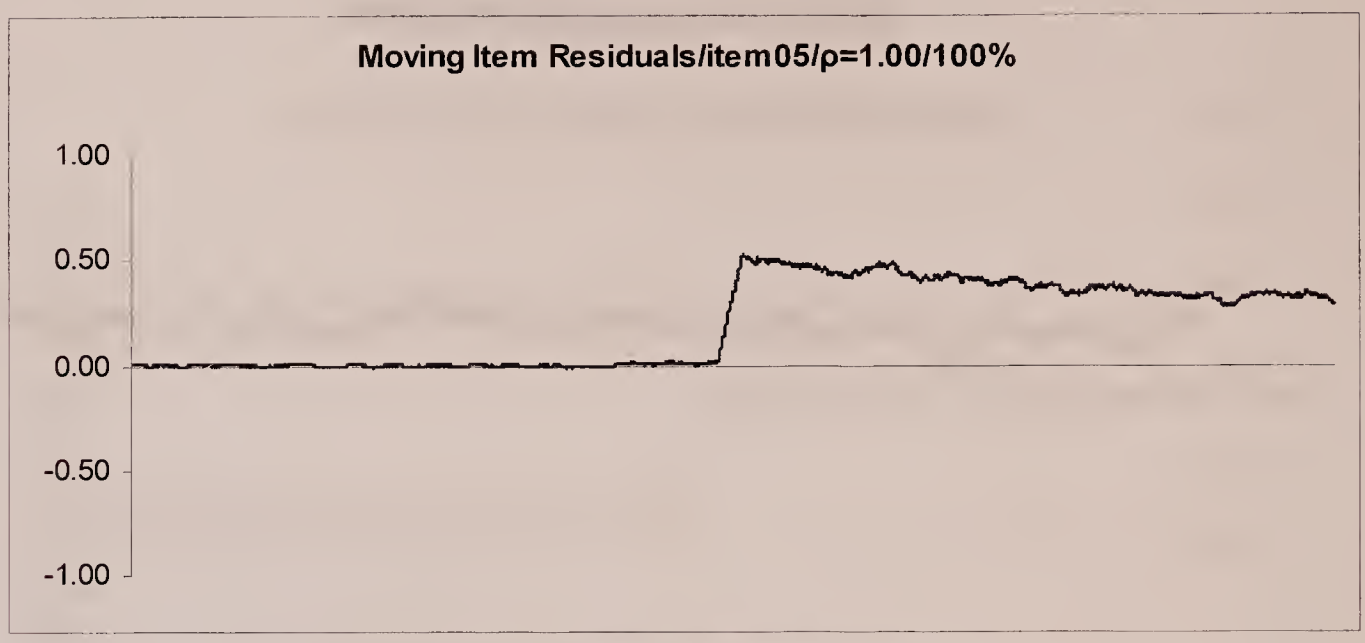


Figure 5.8. plot of item exposure statistics for item 5. (gradually shifting ability distribution, ρ = 1.0, 100%)



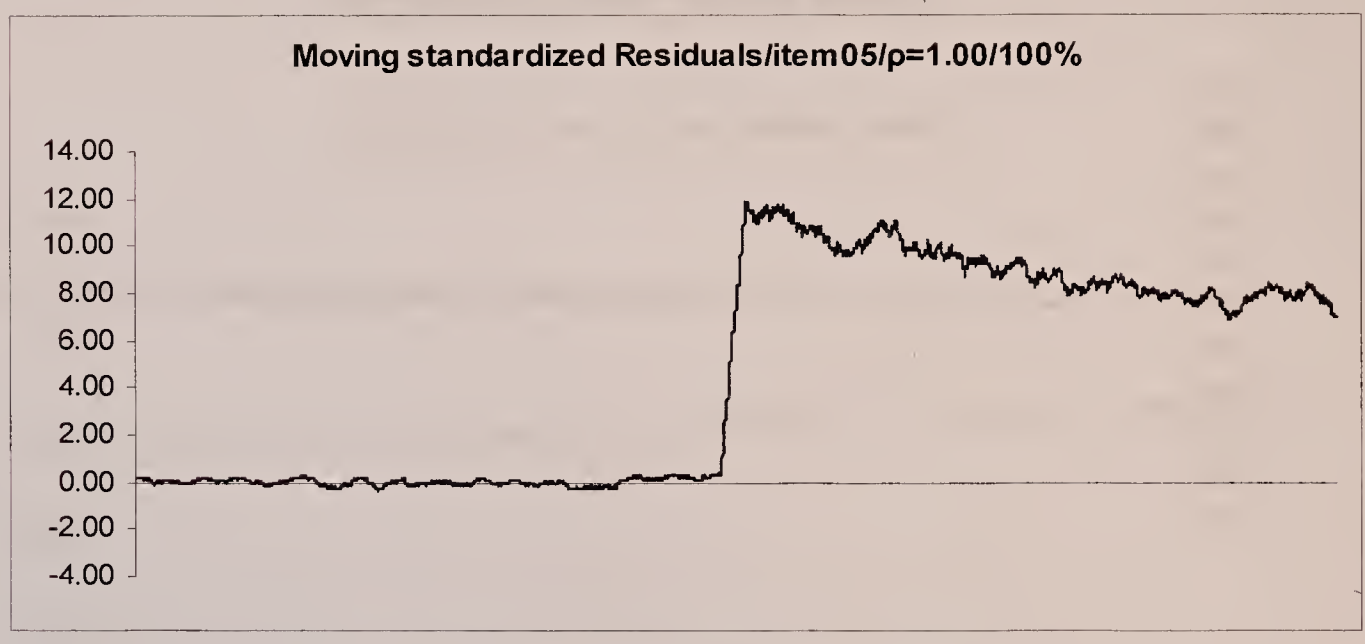
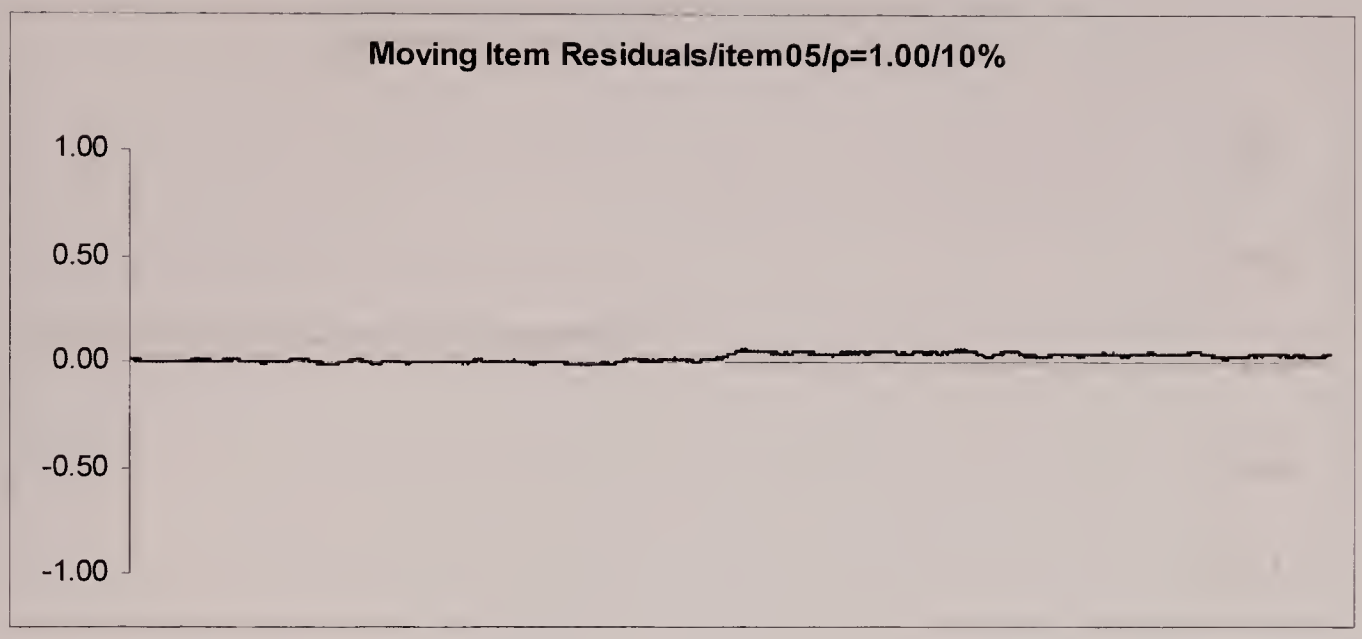


Figure 5.9. plot of item exposure statistics for item 5. (gradually shifting ability distribution, ρ = 1.0, 10%)



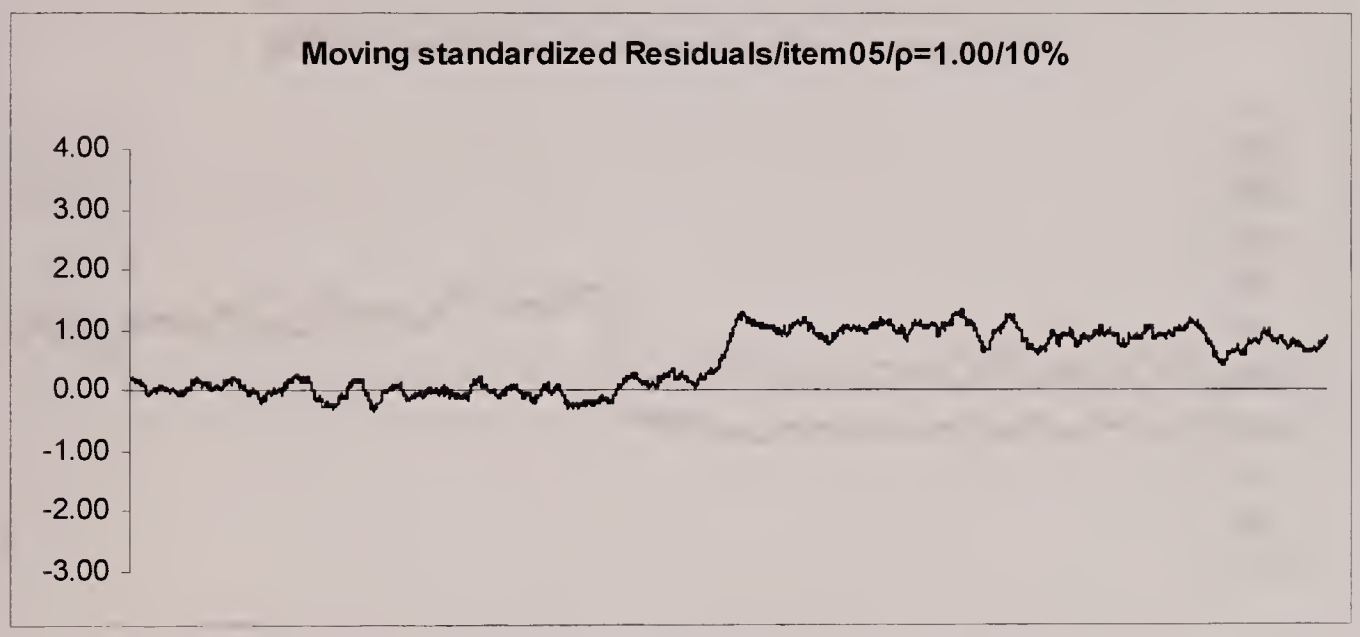
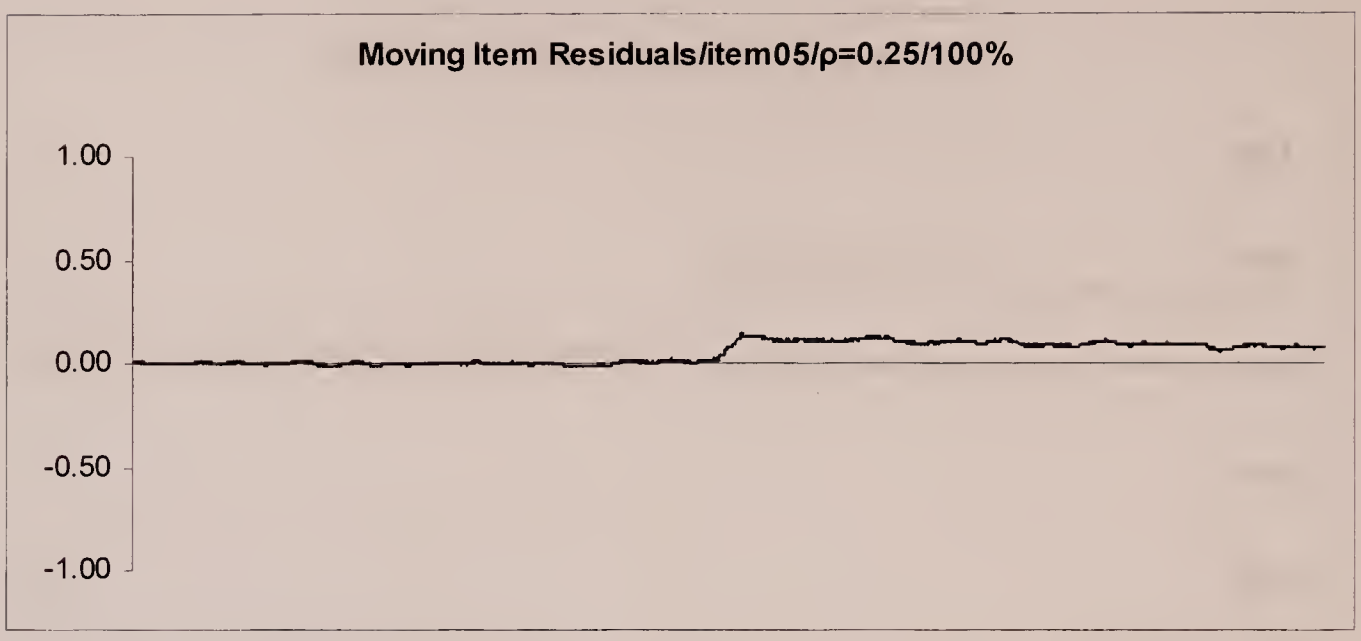


Figure 5.10. plot of item exposure statistics for item 5. (gradually shifting ability distribution, $\rho = 0.25$, 100%)



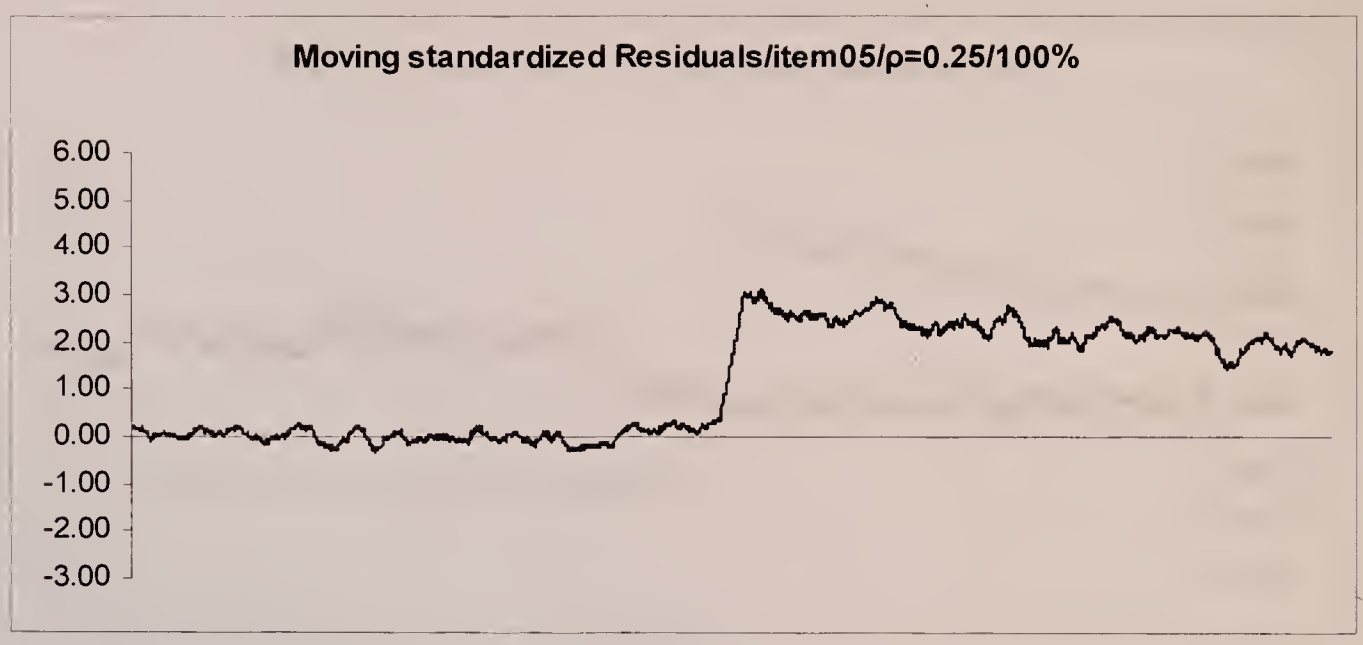
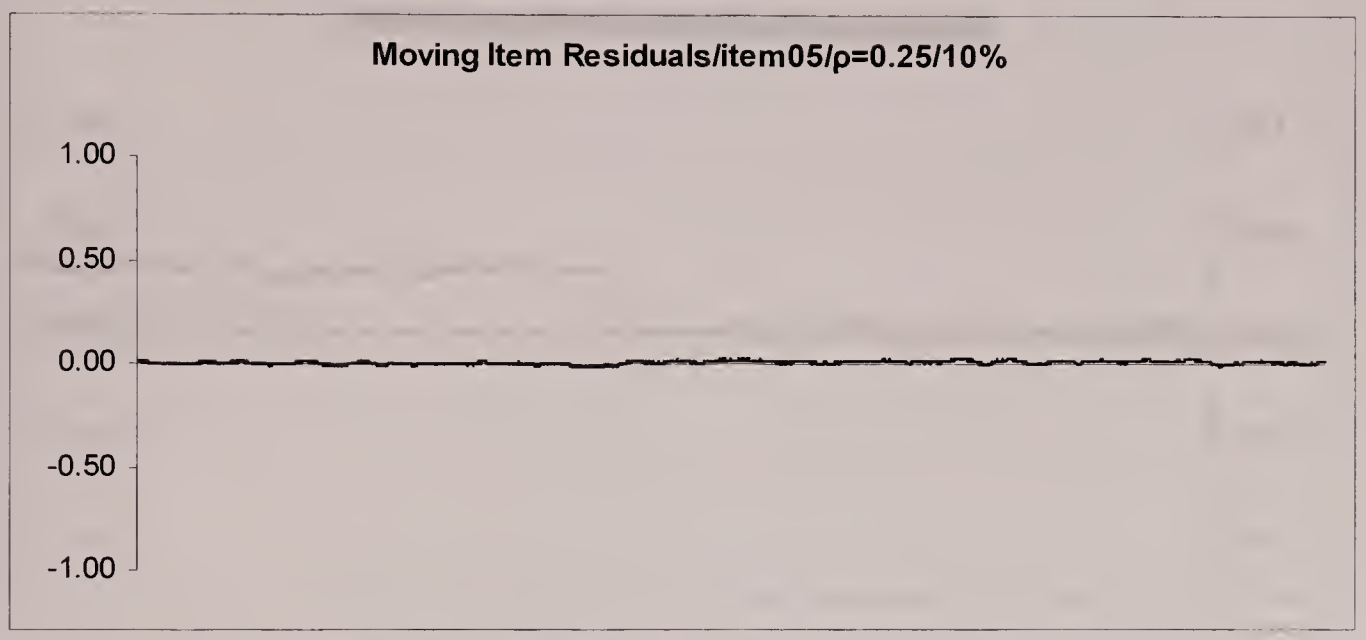


Figure 5.11. plot of item exposure statistics for item 5. (gradually shifting ability distribution, ρ = 0.25, 10%)



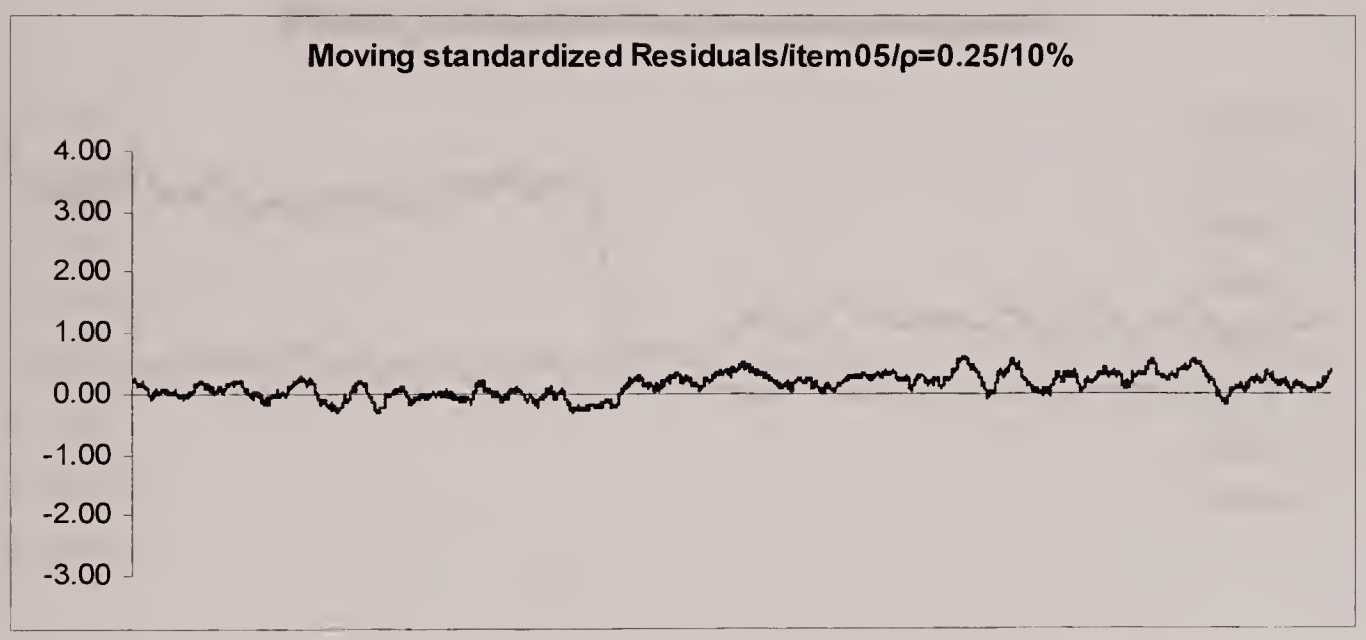
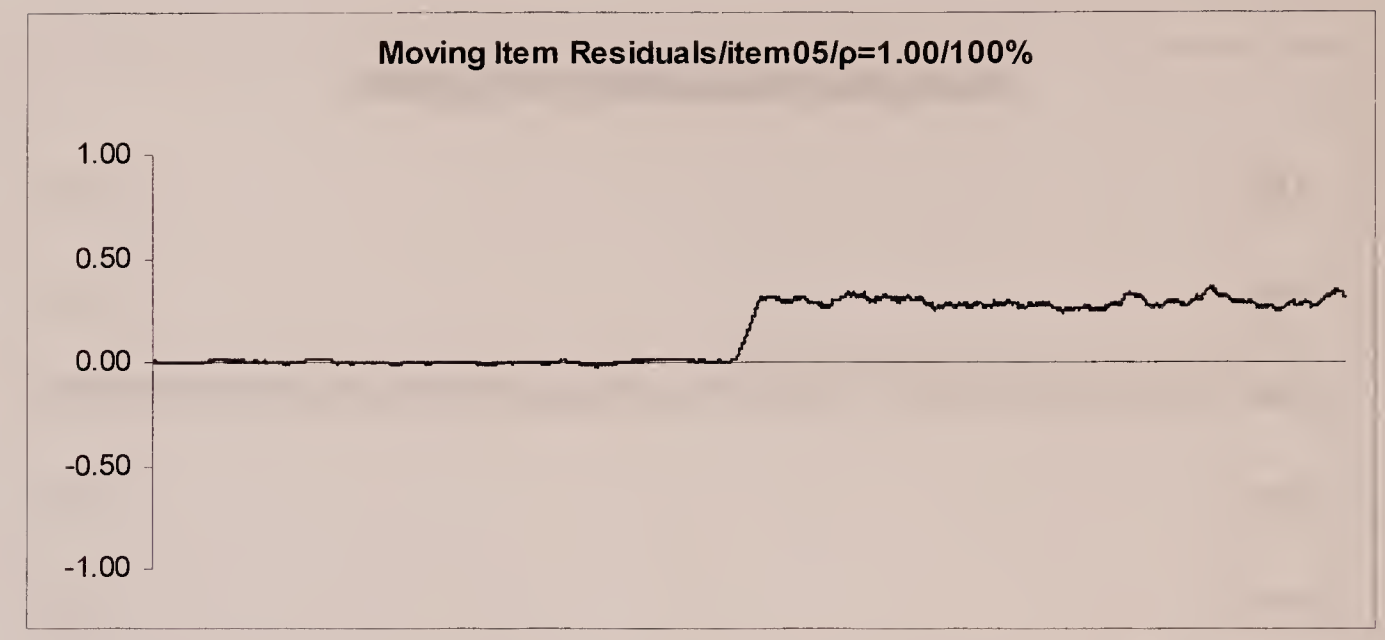


Figure 5.12. plot of item exposure statistics for item 5. (abrupt shifting ability distribution, ρ = 1.0, 100%)



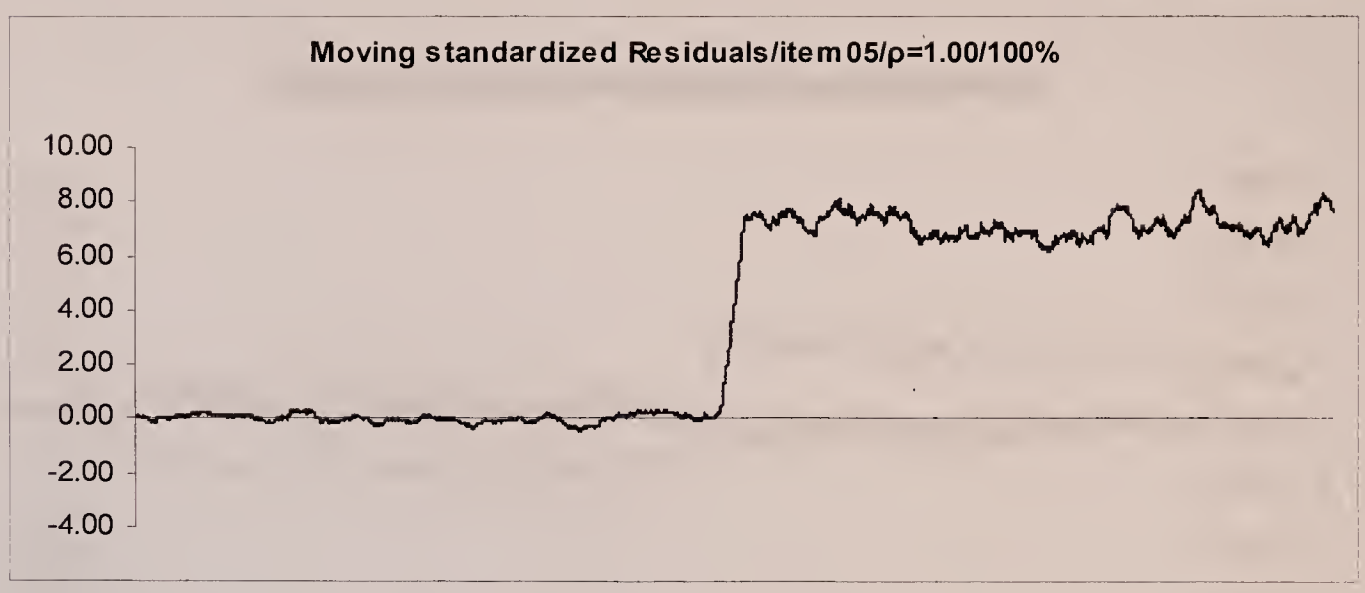
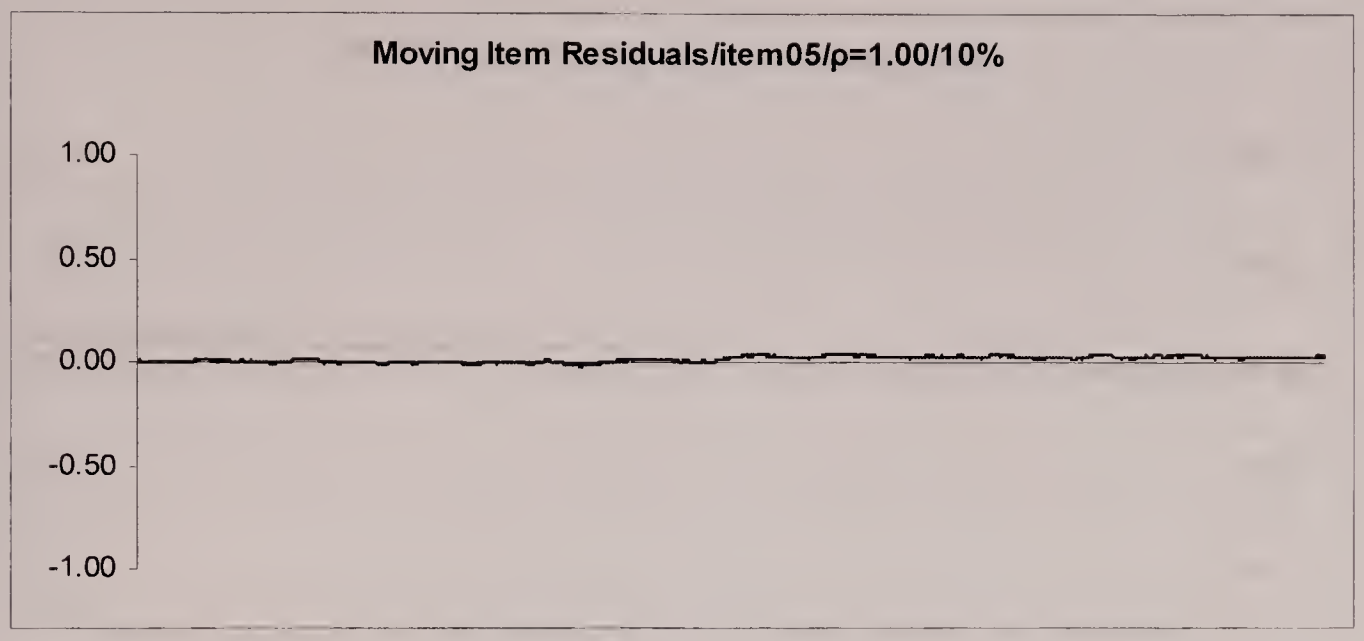


Figure 5.13. plot of item exposure statistics for item 5. (abrupt shifting ability distribution, ρ = 1.0, 10%)



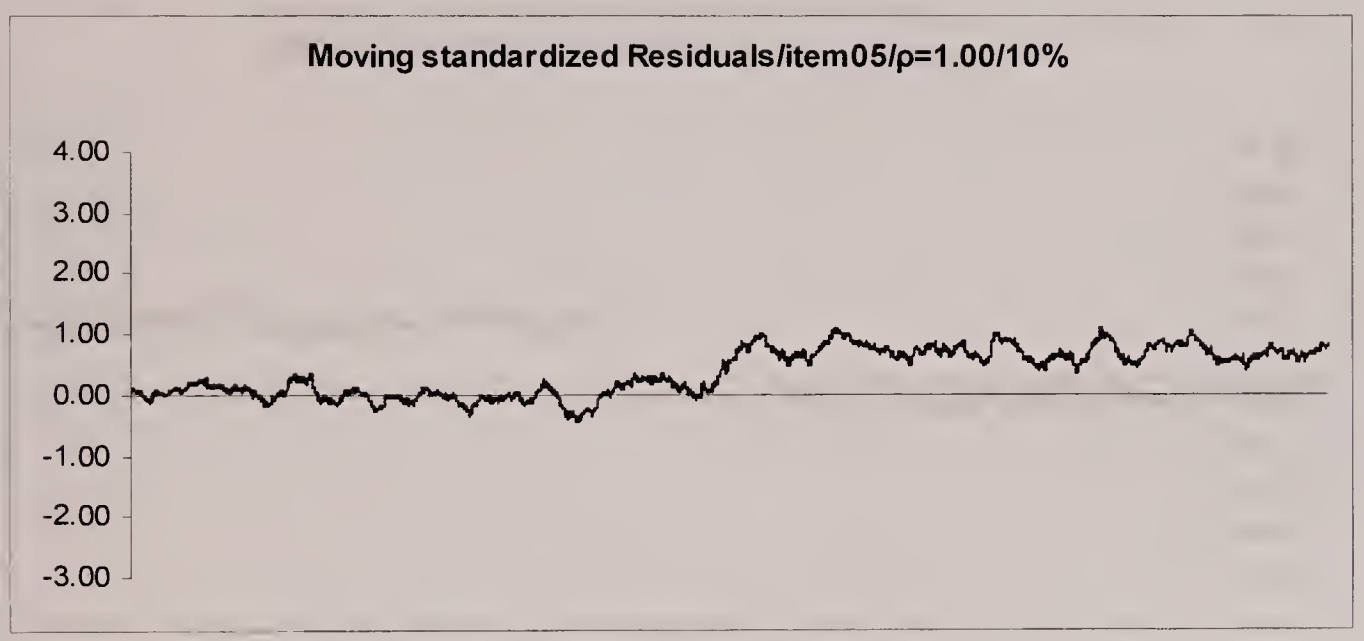
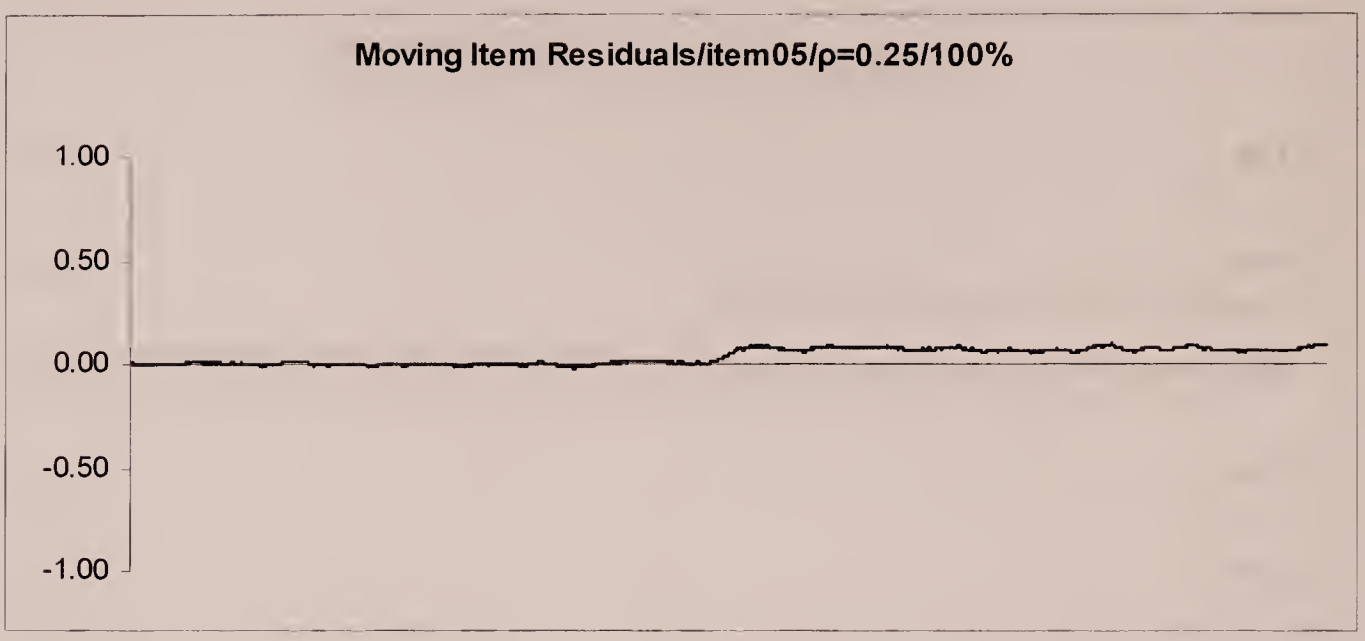


Figure 5.14. plot of item exposure statistics for item 5. (abrupt shifting ability distribution, ρ = 0.25, 100%)



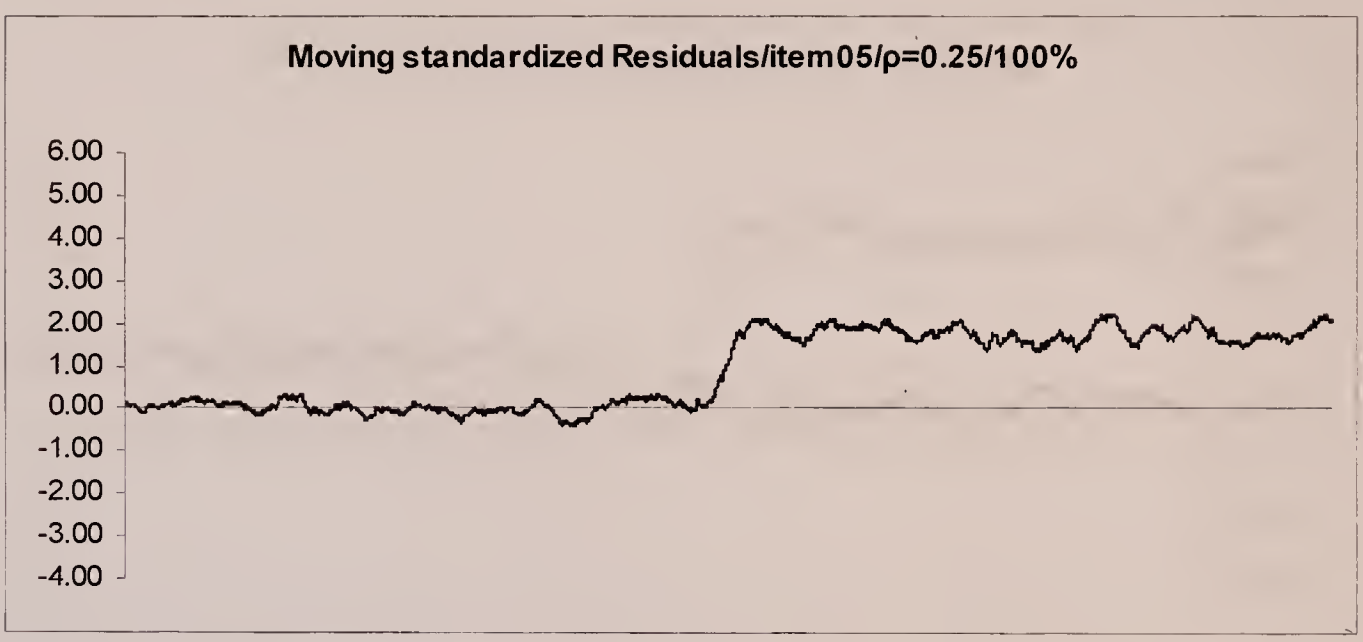
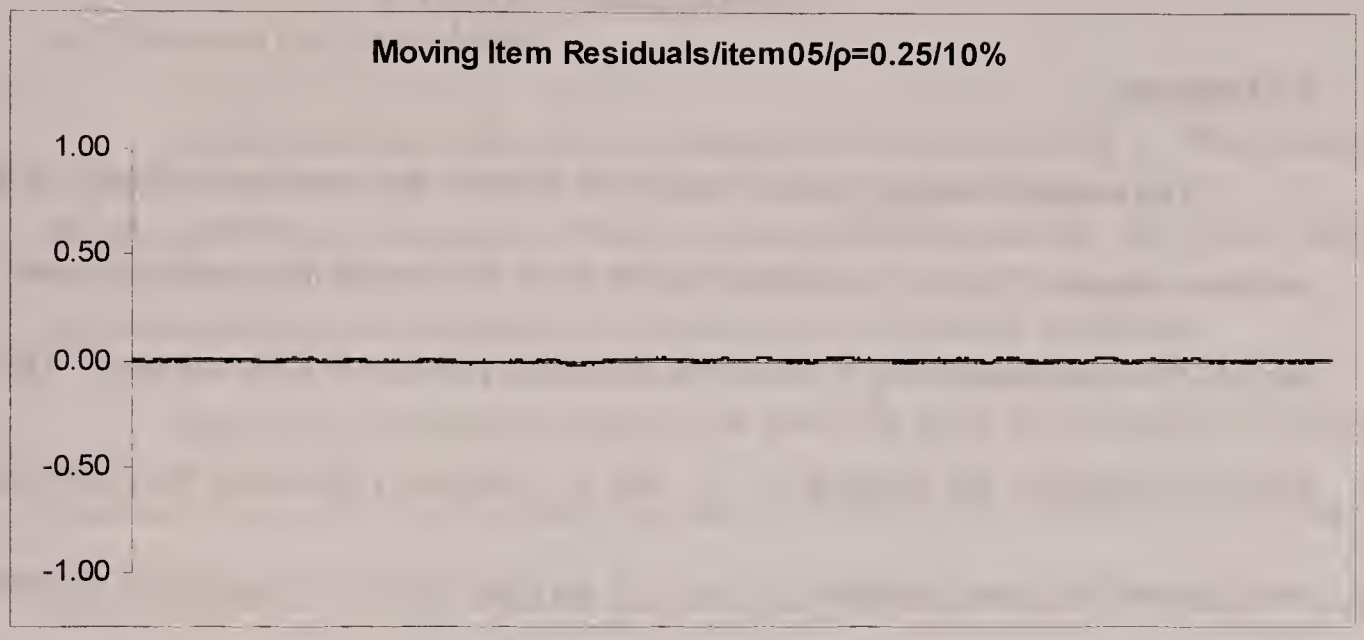
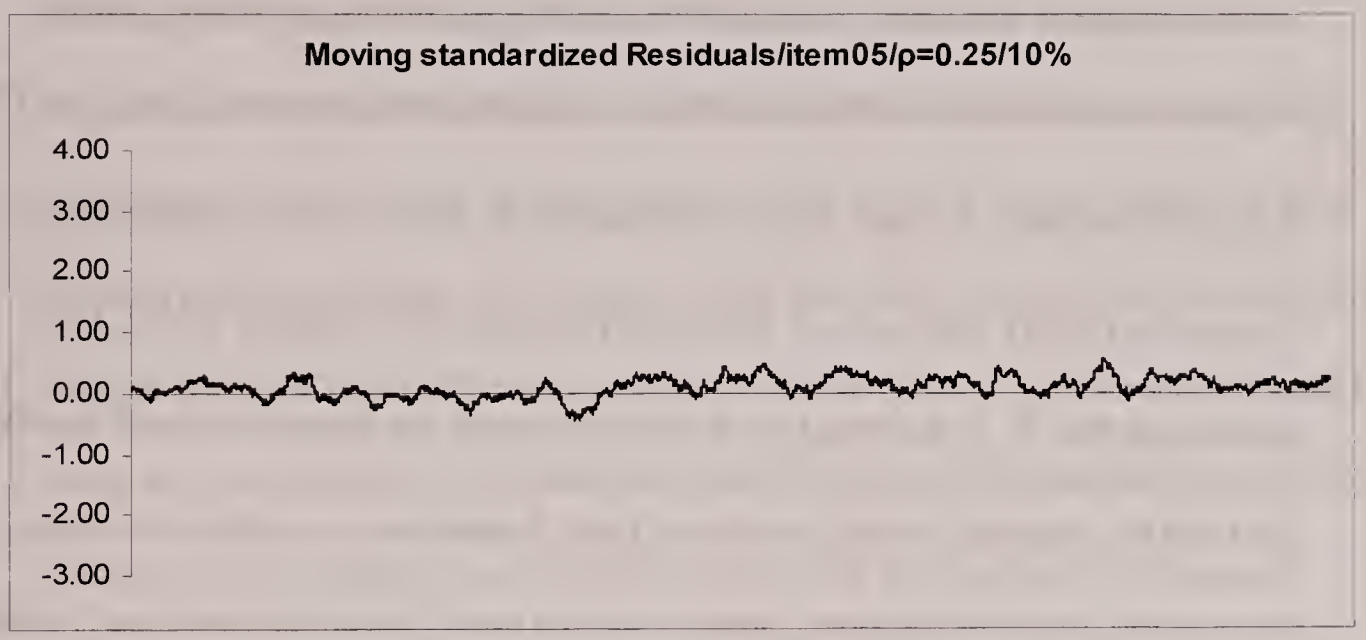


Figure 5.15. plot of item exposure statistics for item 5. (abrupt shifting ability distribution, $\rho = 0.25$, 10%)





CHAPTER 6

SIMULATION STUDY 3

6.1 Purposes

The research design for this simulation study is the same as the design in the previous chapter. The only difference is that three IRT-based item statistics were employed while the moving p value was completely removed from the study. The three item statistics are denoted, Z_c, z_2 and z_3. Generally speaking, they are more or less standardized item residuals. Z_c looks at the total number of examinees obtaining correct responses to an item, and compares it to the expected number of correct responses based on the ability distribution of the examinee group assuming the IRT model fits the data. z_2 and z_3 are extensions of Z_c but z_2 and z_3, according to Zhu, Yu and Liu (2002), improves Z_c while carrying over the simplicity of Z_c. They pointed out that Z_c is directed to a deviation between the observed overall number of right and the expected overall number of right. It measures an average deviation between the observed response function and the model predicted response function across a pre-defined ability range. This index may work well if an item consistently exhibits easier or harder than expected for examinees at all ability levels but performs poorly when an item is differentially harder or easier than expected, conditioned on ability. On the other hand, z_2 and z_3 should perform better in this case.

Therefore, in addition to the classical type of standardized item residual, two more types of standardized statistics were introduced. The purposes of this research were (1) to evaluate three IRT-based standardized residuals in the presence of shifts in the ability distribution over time, (2) to investigate the power and type I error rate of

each detection statistics, and (3) to investigate item exposure detection for items with different statistical characteristics.

6.2 Details of the Methodology

Variables under study were the same as in simulation study 2. They were (1) ability distribution, (2) choice of item exposure detection statistic, (3) type of item exposure model, and (4) statistical characteristics of exposed test items.

The whole simulation process of the previous study was repeated in almost the same manner except that two new item statistics were used. Moreover, the same random number seed was used for programming purposes therefore the exact same item response patterns were generated. Essentially, these two studies can be considered as a whole. The reason to divide them into two pieces was so the first study results were focused on comparisons of the classical item p value, raw item residual and standardized item residual, while this study was focused on a comparison of three different standardizations of item residuals. The only difference between this study and the previous one is that these three IRT-based item statistics were used in all three ability distributions while the item p value was ignored in the case of unstable ability distributions.

Power and type I error rate for each statistic were computed and tabulated again. Unlike the previous study, item detection charts were not displayed in this chapter since the plots for Z_c can be found in Chapter 5 and the other statistics are very similar with Z_c so that only minor differences would be observed.

6.3 Findings

In short, the three item detection statistics performed in very similar ways while z_2 and z_3 showed slight superiority in some circumstances.

Tables 6.1 to 6.8 provide the relevant information about speed of detection, type I error rates and power of detection for items with various statistical properties under different item exposure parameters for the fixed normal ability distribution. The situation with $\rho = 1.0$ and 100% of the examinees benefiting from the exposed information is not interesting at all. All three detecting statistics detected the problem without any difficulty while very few type I errors were committed. When $\rho = 1.0$ and 10% of the examinees benefited from the exposure information, easy items (with b parameter lower than 0) were hard to spot. When $\rho = 0.25$ and 100% of the examinees benefited from the exposed information, the three statistics showed very high power and low type I error rates to most items. For example, they all showed power for all items except the easiest one which was 76.6% while the type I error rates were all lower than 3.0%. For more realistic situations--$\rho = 0.25$ and 10% of the examinees benefiting from the exposed information--the power for all statistics was very low. Importantly, hard and highly discriminated items were still easy to spot. It was noticeable that z_2 and z_3 performed a little bit better than Z_c.

Tables 6.9 to 6.16 provide relevant information for the case with a gradual change in ability from a mean of -1.0 to a mean of +1.0. The pattern was almost the same with the fixed normal distribution. Once again, all statistics spotted all exposed items without any difficulty when $\rho = 1.0$ and 100% of the examinees benefited from the exposed information. When $\rho = 1.0$ and 10% of the examinees benefited from the exposed information, all statistics responded to the trend slower than they did when the ability distribution was fixed. It is necessary and interesting to investigate why this happened in future research. All of the three statistics did not have much difficulty to spot exposed items for the situation that $\rho = 0.25$ and 100% of the examinees

benefiting from the exposed information but they all had trouble with the situation that ρ = 0.25 and 10% of the examinees benefiting from the exposed information.

Tables 6.17 to 6.24 provide relevant information for abrupt change in the mean of the ability distribution. The same patterns as observed in the previous section happened again.

6.4 Conclusions

The simulation results indicated Z_c , z_2 , and z_3 perform almost in the same manner. However, Zhu, Yu and Liu (2002) pointed out that z_2 and z_3 perform better than Z_c when an item is differentially harder or easier than expected, conditioned on ability. This research did not replicate the result due to the limitation of the item exposure simulation model. When an item is exposed it always seems to be easier than it should be so the situation discussed in Zhu, Yu and Liu (2002) may not occur in out context. These three detection indices did show though that for some situations z_2 and z_3 performed a little bit better than Z_c but the evidence is not strong enough to generalize the findings.

Table 6.1. Number of times of item administration after exposure. (ρ = 1.0, for 100%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	24.4	21.3	28.1
	b= 0.00	15.1	10.8	9.5
	b= 1.00	11.7	7.8	4.6
	b= 2.00	9.8	5.0	2.9
Z_3	b=-1.00	24.3	22.2	24.2
	b= 0.00	16.1	10.4	11.3
	b= 1.00	12.1	8.3	7.9
	b= 2.00	10.2	6.2	3.9
Z_c	b=-1.00	25.2	22.6	23.5
	b= 0.00	16.3	12.4	10.9
	b= 1.00	12.4	8.6	7.5
	b= 2.00	10.4	6.7	4.6

Table 6.2. Type I errors and power. (ρ = 1.0, for 100%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	1.58	100.0	3.32	100.0	2.73	100.0
	b= 0.00	2.78	100.0	4.44	100.0	3.40	100.0
	b= 1.00	2.36	100.0	2.89	100.0	5.56	100.0
	b= 2.00	1.89	100.0	3.38	100.0	6.67	100.0
Z_3	b=-1.00	2.13	100.0	2.78	100.0	1.98	100.0
	b= 0.00	2.56	100.0	3.33	100.0	1.94	100.0
	b= 1.00	2.38	100.0	2.59	100.0	2.88	100.0
	b= 2.00	2.01	100.0	2.61	100.0	3.14	100.0
Z_c	b=-1.00	2.15	100.0	2.78	100.0	2.16	100.0
	b= 0.00	2.55	100.0	3.10	100.0	1.94	100.0
	b= 1.00	2.45	100.0	2.74	100.0	2.88	100.0
	b= 2.00	2.09	100.0	2.59	100.0	2.99	100.0

Table 6.3. Number of times of item administration after exposure. ($\rho = 1.0$, for 10%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	302.7	289.2	272.3
	b= 0.00	171.3	134.2	130.7
	b= 1.00	113.3	105.8	41.8
	b= 2.00	91.5	58.8	40.1
Z_3	b=-1.00	253.7	293.3	288.9
	b= 0.00	190.7	133.6	168.8
	b= 1.00	110.4	103.8	56.5
	b= 2.00	88.6	51.6	38.2
Z_c	b=-1.00	283.8	315.4	307.2
	b= 0.00	192.9	149.1	189.5
	b= 1.00	135.0	112.4	67.8
	b= 2.00	96.9	60.8	48.8

Table 6.4. Type I errors and power. ($\rho = 1.0$, for 10%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	1.98	8.8	3.16	11.8	2.03	11.0
	b= 0.00	2.48	19.6	3.41	30.7	3.03	31.2
	b= 1.00	2.17	28.96	2.82	39.5	4.58	69.8
	b= 2.00	1.98	53.4.	2.08	79.9	5.06	96.9
Z_3	b=-1.00	2.13	10.3	2.71	10.8	1.99	11.9
	b= 0.00	2.35	19.8	2.97	26.8	1.93	25.7
	b= 1.00	2.16	37.4	2.44	43.2	2.68	67.6
	b= 2.00	2.01	49.9	2.61	76.6	3.00	96.1
Z_c	b=-1.00	2.15	9.9	2.78	10.0	2.16	9.1
	b= 0.00	2.54	16.7	3.10	23.4	1.94	24.3
	b= 1.00	2.45	29.9	2.74	42.2	2.88	63.5
	b= 2.00	2.08	49.4	2.59	74.9	2.99	93.5

Table 6.5. Number of times of item administration after exposure. ($\rho = 0.25$, for 100%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	100.5	103.5	101.8
	b= 0.00	60.6	50.3	65.2
	b= 1.00	51.1	44.2	22.7
	b= 2.00	33.6	22.1	11.6
Z_3	b=-1.00	99.9	109.8	100.9
	b= 0.00	64.4	51.8	51.1
	b= 1.00	49.1	41.3	29.1
	b= 2.00	31.8	22.2	17.8
Z_c	b=-1.00	99.3	119.1	109.3
	b= 0.00	64.9	52.9	56.0
	b= 1.00	46.3	45.6	29.7
	b= 2.00	38.3	25.1	20.8

Table 6.6. Type I errors and power. ($\rho = 0.25$, for 100%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	1.52	41.1	3.33	43.9	2.33	48.3
	b= 0.00	2.48	76.6	4.12	79.9	2.63	88.7
	b= 1.00	2.14	88.6	2.56	99.7	2.55	99.8
	b= 2.00	1.98	100.0	3.48	100.0	2.06	100.0
Z_3	b=-1.00	2.04	43.7	2.77	39.9	1.97	47.1
	b= 0.00	2.25	74.3	3.08	83.8	1.96	89.5
	b= 1.00	2.29	89.8	2.46	98.8	2.94	99.9
	b= 2.00	2.01	99.3	2.52	100.0	2.71	100.0
Z_c	b=-1.00	2.15	47.2	2.78	39.8	2.16	42.0
	b= 0.00	2.54	74.0	3.10	80.5	1.94	89.2
	b= 1.00	2.45	91.4	2.74	98.3	2.88	99.8
	b= 2.00	2.08	99.4	2.59	100.0	2.99	100.0

Table 6.7. Number of times of item administration after exposure. ($\rho = 0.25$, for 10%, normal distribution of ability)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	617.5	573.4	693.3
	b= 0.00	499.9	420.4	390.2
	b= 1.00	533.9	344.0	198.8
	b= 2.00	427.7	193.7	146.6
Z_3	b=-1.00	626.6	602.4	701.1
	b= 0.00	462.8	538.2	478.8
	b= 1.00	538.3	345.4	271.1
	b= 2.00	470.8	241.6	179.5
Z_c	b=-1.00	650.5	671.9	674.7
	b= 0.00	482.9	591.6	479.0
	b= 1.00	573.2	388.0	282.3
	b= 2.00	474.4	255.3	180.5

Table 6.8. Type I errors and power. ($\rho = 0.25$, for 10%, normal distribution of ability)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	2.05	4.14	3.36	4.0	2.00	4.2
	b= 0.00	2.28	4.59	4.41	5.6	2.03	6.6
	b= 1.00	2.36	5.44	2.86	7.5	2.85	13.1
	b= 2.00	1.99	6.82	4.08	11.1	3.01	25.7
Z_3	b=-1.00	2.15	3.97	2.48	3.5	1.91	4.4
	b= 0.00	2.45	5.68	3.07	6.0	1.88	5.8
	b= 1.00	2.33	6.88	2.66	7.3	2.81	10.9
	b= 2.00	2.03	7.73	2.37	10.8	3.00	20.1
Z_c	b=-1.00	2.15	3.78	2.78	3.4	2.16	3.6
	b= 0.00	2.54	4.67	3.10	5.8	1.94	4.9
	b= 1.00	2.45	6.23	2.74	7.8	2.88	10.6
	b= 2.00	2.08	7.24	2.59	11.5	2.99	20.6

Table 6.9. Number of times of item administration after exposure. (ρ = 1.0, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	24.1	28.5	14.8
	b= 0.00	15.9	14.0	12.9
	b= 1.00	12.1	9.9	8.8
	b= 2.00	9.8	8.6	3.9
Z_3	b=-1.00	23.8	23.4	26.6
	b= 0.00	17.4	13.8	11.8
	b= 1.00	12.4	11.1	9.4
	b= 2.00	10.6	9.8	4.0
Z_c	b=-1.00	23.2	22.8	26.7
	b= 0.00	16.8	13.4	12.1
	b= 1.00	13.1	10.4	9.0
	b= 2.00	10.9	7.8	4.3

Table 6.10. Type I errors and power. (ρ = 1.0, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	2.9	100.0	1.7	100.0	2.5	100.0
	b= 0.00	3.6	100.0	2.4	100.0	2.9	100.0
	b= 1.00	2.2	100.0	2.1	100.0	2.0	100.0
	b= 2.00	2.1	100.0	2.3	100.0	3.2	100.0
Z_3	b=-1.00	3.1	100.0	1.7	100.0	2.3	100.0
	b= 0.00	3.4	100.0	2.9	100.0	2.6	100.0
	b= 1.00	2.6	100.0	1.9	100.0	2.4	100.0
	b= 2.00	2.2	100.0	1.9	100.0	3.2	100.0
Z_c	b=-1.00	3.1	100.0	1.5	100.0	2.4	100.0
	b= 0.00	3.2	100.0	2.8	100.0	2.7	100.0
	b= 1.00	2.5	100.0	2.2	100.0	2.5	100.0
	b= 2.00	2.3	100.0	2.0	100.0	3.6	100.0

Table 6.11. Number of times of item administration after exposure. (ρ = 1.0, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	242.4	188.1	324.4
	b= 0.00	181.5	170.3	151.9
	b= 1.00	140.1	95.1	72.2
	b= 2.00	110.8	79.9	47.8
Z_3	b=-1.00	285.9	302.0	298.9
	b= 0.00	182.1	173.6	123.9
	b= 1.00	131.7	88.8	77.5
	b= 2.00	116.3	74.4	47.8
Z_c	b=-1.00	280.0	304.4	320.9
	b= 0.00	190.0	170.5	149.1
	b= 1.00	134.8	89.4	74.1
	b= 2.00	114.2	73.7	47.9

Table 6.12. Type I errors and power. (ρ = 1.0, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	2.9	6.4	1.7	17.0	2.5	7.3
	b= 0.00	3.3	12.3	3.0	17.1	2.8	13.3
	b= 1.00	2.2	21.7	2.1	27.7	2.5	39.6
	b= 2.00	2.1	42.5	2.0	57.3	3.2	79.1
Z_3	b=-1.00	3.0	5.3	1.6	7.8	2.4	5.4
	b= 0.00	3.4	14.7	2.9	16.4	2.8	12.8
	b= 1.00	2.8	28.0	1.9	30.1	2.4	38.3
	b= 2.00	2.2	44.6	1.9	56.7	3.2	76.9
Z_c	b=-1.00	3.1	7.1	1.5	6.9	2.4	6.1
	b= 0.00	3.2	12.1	2.8	15.2	2.7	12.9
	b= 1.00	2.5	19.9	2.2	25.3	2.5	37.2
	b= 2.00	2.3	38.0	2.0	54.4	3.6	74.7

Table 6.13. Number of times of item administration after exposure. ($\rho = 0.25$, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	90.3	126.5	117.3
	b= 0.00	60.7	67.6	48.4
	b= 1.00	47.3	45.2	31.8
	b= 2.00	44.5	30.2	15.9
Z_3	b=-1.00	88.9	122.9	128.8
	b= 0.00	60.9	55.8	52.0
	b= 1.00	45.7	42.2	29.7
	b= 2.00	45.1	27.7	17.2
Z_c	b=-1.00	90.1	124.3	128.8
	b= 0.00	62.2	56.0	54.2
	b= 1.00	46.8	42.3	30.2
	b= 2.00	45.6	28.4	17.7

Table 6.14. Type I errors and power. ($\rho = 0.25$, for 100%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	3.0	29.8	1.8	18.8	2.5	19.0
	b= 0.00	3.1	57.4	3.0	60.1	2.9	56.8
	b= 1.00	2.4	83.2	2.9	90.0	2.3	97.4
	b= 2.00	2.1	97.0	2.1	99.9	3.2	100.0
Z_3	b=-1.00	3.3	29.6	1.7	17.2	2.4	17.3
	b= 0.00	3.3	58.9	2.9	57.3	2.8	56.9
	b= 1.00	2.6	85.5	2.1	92.3	2.4	99.0
	b= 2.00	2.2	98.4	1.9	99.9	3.4	100.0
Z_c	b=-1.00	3.1	33.7	1.5	19.5	2.4	16.3
	b= 0.00	3.2	58.0	2.8	57.7	2.7	57.4
	b= 1.00	2.5	81.3	2.2	89.4	2.5	95.7
	b= 2.00	2.3	96.8	2.0	99.5	3.6	100.0

Table 6.15. Number of times of item administration after exposure. ($\rho = 0.25$, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	613.3	514.2	480.7
	b= 0.00	528.1	491.2	351.5
	b= 1.00	452.9	484.4	290.8
	b= 2.00	393.6	295.5	177.1
Z_3	b=-1.00	596.8	501.9	497.1
	b= 0.00	507.7	496.1	378.4
	b= 1.00	459.4	482.6	283.2
	b= 2.00	394.6	282.9	165.9
Z_c	b=-1.00	609.8	528.9	518.7
	b= 0.00	548.5	497.7	375.9
	b= 1.00	477.2	493.0	296.5
	b= 2.00	409.6	306.9	177.6

Table16. Type I errors and power. ($\rho = 0.25$, for 10%, gradual change in ability from a mean of -1.0 to a mean of +1.0)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	2.9	2.7	1.7	4.8	2.5	3.5
	b= 0.00	3.0	2.9	2.9	5.9	2.7	5.2
	b= 1.00	2.2	5.4	2.1	6.3	2.5	16.3
	b= 2.00	2.1	9.2	1.8	7.2	3.2	15.8
Z_3	b=-1.00	3.1	2.9	1.3	3.3	2.4	3.1
	b= 0.00	3.1	3.1	2.9	4.5	2.7	4.9
	b= 1.00	2.5	5.6	2.1	6.7	2.4	11.7
	b= 2.00	2.2	9.9	1.9	7.1	3.2	13.6
Z_c	b=-1.00	3.1	2.7	1.5	2.6	2.4	3.1
	b= 0.00	3.2	2.8	2.8	4.9	2.7	4.4
	b= 1.00	2.5	4.1	2.2	5.8	2.5	6.9
	b= 2.00	2.3	6.4	2.0	6.5	3.6	12.3

Table 6.17. Number of times of item administration after exposure. (ρ = 1.0, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	41.7	54.2	56.9
	b= 0.00	23.8	30.1	24.5
	b= 1.00	18.0	20.4	9.3
	b= 2.00	11.1	8.4	3.0
Z_3	b=-1.00	39.9	58.2	61.3
	b= 0.00	23.6	22.3	24.3
	b= 1.00	15.1	14.7	9.1
	b= 2.00	11.7	5.1	2.9
Z_c	b=-1.00	40.6	50.5	70.0
	b= 0.00	23.5	22.3	25.3
	b= 1.00	15.0	12.5	8.0
	b= 2.00	11.2	5.3	2.2

Table 6.18. Type I errors and power. (ρ = 1.0, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	3.0	100.0	1.7	100.0	2.4	100.0
	b= 0.00	2.4	100.0	3.9	100.0	2.4	100.0
	b= 1.00	2.0	100.0	3.3	100.0	2.9	100.0
	b= 2.00	1.9	100.0	3.4	100.0	3.9	100.0
Z_3	b=-1.00	3.1	100.0	1.8	100.0	2.3	87.8
	b= 0.00	2.2	100.0	3.9	100.0	2.4	100.0
	b= 1.00	1.9	100.0	3.3	100.0	2.5	100.0
	b= 2.00	1.9	100.0	3.4	100.0	3.7	100.0
Z_c	b=-1.00	2.7	100.0	1.7	100.0	2.2	100.0
	b= 0.00	2.3	100.0	3.9	100.0	3.3	100.0
	b= 1.00	2.6	100.0	3.3	100.0	2.9	100.0
	b= 2.00	2.0	100.0	3.5	100.0	4.2	100.0

Table 6.19. Number of times of item administration after exposure. (ρ = 1.0, for 10%, abrupt change in the mean of the ability distribution)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	445.2	583.4	695.3
	b= 0.00	261.0	216.8	277.7
	b= 1.00	172.7	108.1	99.6
	b= 2.00	141.4	82.7	55.4
Z_3	b=-1.00	425.9	572.3	715.0
	b= 0.00	249.6	206.1	274.5
	b= 1.00	192.9	118.6	99.1
	b= 2.00	131.4	82.2	55.4
Z_c	b=-1.00	413.9	601.8	669.4
	b= 0.00	248.9	236.2	274.2
	b= 1.00	206.5	194.3	116.0
	b= 2.00	149.9	95.0	57.4

Table 6.20. Type I errors and power. (ρ = 1.0, for 10%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	3.7	3.7	1.7	4.8	2.2	4.7
	b= 0.00	2.2	9.4	3.9	10.0	3.1	7.4
	b= 1.00	2.2	17.2	3.2	19.8	2.7	22.8
	b= 2.00	1.9	35.1	3.3	39.7	4.4	61.7
Z_3	b=-1.00	3.7	4.4	1.6	4.5	2.3	4.6
	b= 0.00	2.2	9.4	3.9	9.9	3.2	7.1
	b= 1.00	2.2	14.1	3.3	18.7	2.5	32.3
	b= 2.00	1.9	33.9	3.4	41.1	4.0	62.2
Z_c	b=-1.00	2.7	6.4	1.7	4.5	2.2	3.4
	b= 0.00	2.3	9.4	3.9	9.9	3.3	7.6
	b= 1.00	2.6	14.7	3.3	19.7	2.9	26.5
	b= 2.00	2.0	32.5	3.5	43.7	4.2	60.6

Table 6.21. Number of times of item administration after exposure. ($\rho = 0.25$, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	187.3	243.6	315.8
	b= 0.00	104.4	83.8	102.1
	b= 1.00	52.6	61.0	40.2
	b= 2.00	50.8	27.1	14.4
Z_3	b=-1.00	183.9	233.4	325.2
	b= 0.00	105.6	88.2	102.2
	b= 1.00	55.2	60.3	40.1
	b= 2.00	50.4	30.1	14.0
Z_c	b=-1.00	173.4	234.0	332.5
	b= 0.00	104.7	89.3	103.3
	b= 1.00	57.5	64.9	38.8
	b= 2.00	50.9	26.6	17.7

Table 6.22. Type I errors and power. ($\rho = 0.25$, for 100%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	2.8	21.1	3.0	14.1	2.3	5.4
	b= 0.00	2.2	44.7	3.7	40.5	3.2	38.6
	b= 1.00	2.6	62.3	3.1	82.0	3.3	94.3
	b= 2.00	2.4	97.6	3.1	101.1	4.1	100.0
Z_3	b=-1.00	3.1	21.8	3.0	16.1	2.3	6.4
	b= 0.00	2.2	44.7	3.9	40.4	3.1	38.5
	b= 1.00	2.4	66.2	3.3	80.5	3.3	91.7
	b= 2.00	1.9	91.6	3.2	96.9	4.2	100.0
Z_c	b=-1.00	2.7	25.6	1.7	15.6	2.2	8.7
	b= 0.00	2.3	45.8	3.9	41.7	3.3	37.1
	b= 1.00	2.6	69.5	3.3	79.7	2.9	86.8
	b= 2.00	2.0	93.9	3.5	97.8	4.2	100.0

Table 6.23. Number of times of item administration after exposure. ($\rho = 0.25$, for 10%, abrupt change in the mean of the ability distribution)

		a=0.40	a=0.70	a=1.20
Z_2	b=-1.00	740.9	634.6	624.2
	b= 0.00	514.4	464.2	543.1
	b= 1.00	468.9	427.4	299.3
	b= 2.00	492.4	390.3	276.6
Z_3	b=-1.00	770.4	636.8	622.6
	b= 0.00	519.5	468.2	541.9
	b= 1.00	488.8	457.9	289.5
	b= 2.00	492.4	340.2	263.7
Z_c	b=-1.00	768.7	719.0	659.4
	b= 0.00	525.1	466.9	556.5
	b= 1.00	505.8	521.8	338.9
	b= 2.00	539.9	407.7	328.7

Table 6.24. Type I errors and power. ($\rho = 0.25$, for 10%, abrupt change in the mean of the ability distribution)

		a=0.40		a=0.70		a=1.20	
		I	II	I	II	I	II
Z_2	b=-1.00	2.6	3.1	1.9	2.4	2.5	2.6
	b= 0.00	2.3	3.1	3.7	3.7	3.0	3.3
	b= 1.00	2.5	4.3	2.7	6.1	2.3	6.1
	b= 2.00	2.0	6.3	3.4	6.4	3.1	7.6
Z_3	b=-1.00	2.8	3.3	2.0	2.5	2.3	1.8
	b= 0.00	2.2	3.4	3.5	3.6	3.2	3.2
	b= 1.00	2.9	4.4	2.8	6.2	2.5	6.2
	b= 2.00	1.9	6.9	3.2	5.9	3.3	6.7
Z_c	b=-1.00	2.7	3.2	1.7	2.2	2.2	2.0
	b= 0.00	2.3	3.3	3.9	3.1	3.3	2.7
	b= 1.00	2.6	3.5	3.3	5.2	2.9	5.4
	b= 2.00	2.0	6.4	3.5	7.2	4.2	8.3

CHAPTER 7

CONCLUSIONS

7.1 Main Findings

The general purpose of this research was to extend the item exposure detection method proposed by Han (2003). In addition to classical item difficulty discussed in Han (2003), the emphasis in this study was focused on a number of IRT-based item statistics. Two sets of IRT-based item exposure detection statistics were investigated in this study. All IRT-based item statistics investigated showed the valuable property that they were free of the underlying candidate ability distribution. That is to say, the moving average sequences of these item statistics were stable over time when the test was administrated under normal situations, that is when ability shifts in the distribution took place. This is a very important property, which will be extremely helpful in practice.

Many other factors can cause the item characteristics drifting and to discover the trend in a timely manner is an important step to maintain the fairness and validity of tests. By monitoring these IRT-based statistics using statistical control charts, tests administered can avoid the complexity of item parameter recalibration. In this study, item residuals and three different types of standardized residuals performed very similarly in terms of speed of detection, power and type I error. The results obtained and reported in Chapters 4 to 6 do not support a simple decision of choice from these statistics. However, standardized residuals showed some superiority over residuals. The theoretical or at least approximate null distribution for standardized residuals seems to be easier to derive so that the determination of control limits makes more sense than that for item residuals.

Although IRT statistics showed the invariant property to different ability distributions, classical item difficulty is also promising for some particular situations. One primary advantage of classical item difficulty is it is model free and computations are easy. Monitoring classical item difficulty can provide meaningful information in timely manner with a low cost. Moreover, if we look at the problem from the other facet, moving p values provide information about how the examinees abilities vary in the testing window. This type of information may be valuable for management purposes.

An unsurprising yet important finding is that the detection capability of any item statistic is closely correlated with the statistics of the items to be monitored. General speaking, hard items are easy to spot and easy items are hard to spot. This provides extra evidence that the proposed method is effective since this conclusion can be easily derived and explained by logic and our experiences. This information is also inspiring to test administers since items with excellent statistics in their pools usually play critical role in testing administration and are more expensive to develop.

Window size is a critical factor to the success of the detection. If a window size is too small, the moving average sequence cannot be stable. On the other hand, a big window size makes the moving average sequence response slow to detect the trend underlying the data. The result of simulation 1 shows that a window size no less than 200 is needed to obtain a stable moving average sequence. This finding is promising from a practical point of view. Several hundred candidates appear to be an ideal number for most educational testing programs.

The statistical control chart is a widely used tool in manufacturing industry but is seldom used in educational and psychological testing. This research strongly recommends the usage of statistical control chart in computer-based testing (CBT).

In addition to the proposed item statistics, some other statistics are also possible too. For example, candidate time information on items is being routinely compiled with many CBT programs. Were candidates to answer an item correctly using substantially less time than other candidates, a question could be raised about the validity of the candidate's response. Possibly, this information can be combined with the item detection statistic to more rapidly identify exposed items. For example, if we monitor the answering time that each examinee spends on one particular item, it may also provide information relevant with item compromising. For more on this, see some of the promising research of Professor van der Linden at the University of Twente in the Netherlands.

In conventional paper and pencil tests, a great number of examinees take the same test at the same time. The scores obtained are random samples of observations. In CBT, the examinees take the test successively. The scores obtained are a sequence of measurements that may follow non-random orders. In statistical terms, they are time series data. There may be some internal structure underlying the data to be interpreted. Moving average is a powerful tool to unveil the structure. When we assume the stream of the examinees over time is equivalent, the variation of the moving averages indicates if an item pool is compromised or an item is exposed. Although the proposed method cannot prevent the item pool from being compromised, it does alert test administrators to the problem so that they can rotate an item pool or remove an item so that fairness and validity are maintained. Technically, it is not expensive to integrate some moving average computation into test administration systems to provide on-site monitoring. The AICPA has already implemented some simple procedures (p-values and item residuals) based on the initial research (Han, 2003).

As pointed out in Chapter 2, although there are not many research studies that directly focus on security issues, many topics can be found to be closely related with it that are being studied. The research direction suggested by this dissertation tried to deal with the security issue from the other facet. New understanding about issues such as what represents acceptable item exposure rates and how long an item pool should be used are important. But, once the test administrators are able to spot exposed items as soon as possible, they are quite aware about what exposure rate is desirable and how long an item pool should be used. This research direction will be complementary with other research directions that try to prevent item pools from being compromised and provide test administrators with a tool to collect validity evidence about CBT scores.

7.2 Future Research

A CAT administration normally requires a large supply of items with accurately estimated psychometric properties in order to sustain continuous testing. However, a pre-calibrated set of test items that is normally obtained during the process of pretest data analysis doesn't always correctly capture what underlies a new set of examinee responses to the item. This so-called model-data deviation can be blamed, at least in part, on the disclosure of the items. In practice, many other factors can contribute to the deviation, such as not perfect initial pretest calibration due to estimation methodology or limited calibration sample size, differences in motivation of the test takers between the pretest and on-line stage, changes in examinees' learning experience, and so on. How to decompose different factors from a mixed trend remains a major challenge to study.

Test and items compromise may be caused by different reasons so it is a challenge to simulate how the examinees' performance will change when an item is

exposed. The item exposure simulation model developed in this study tried to vary the extent of item compromise by varying one parameter. From the point of view of the study, this simulation model is very close to the situation that was described as "organizational theft" in Change and Zhang (2002), which has been considered as a major threaten to CBT according to some incidents that happened in Asia. However, the time factor seems to be important as well to compromise an item. The longer an item pool has been used, the more items are possibly known by examinees. This feature should be taken into account in a future study.

A characteristic of CAT is that the range of the examinees' abilities will become very narrow in test administration. The item selection algorithm of CAT usually chooses items whose difficulty parameters are close to the ability estimates of examinees. Therefore, most items in CBT are not administered to examinees from a wide ability range but from a narrow range. The findings of this study need to be tempered by the ability range of the examinees.

Moreover, determination of control limits seems to be a problem for the simulation studies since they played a critical role in the computation of power and type I error. However, it may not so important in practice since it may become a judgment problem. The bigger the upper limits are, the lower power the method has and fewer type I errors are made, but the more questionable the detection capability becomes. On the other hand, if the upper limits are set up too low, more false alarms will be sounded. How the detection statistics perform when the control limits vary should be investigated.

Currently, different item delivery algorithms have been employed from fixed linear, multi-stage testing, to CAT. Although it has been pointed out that the proposed item exposure detection process is free of the item delivery mechanism,

some item delivery methods pack items into different testlets or blocks and the items within each testlet or block cannot be replaced individually. How to deal with an exposed item remains a problem even though it was detected. Further research is required on how to block or eliminate the exposed items for multi-stage designs.

BIBLIOGRAPHY

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Assessment Systems Corporation. (1993). *Scrutiny: software to identify test misconduct.* St. Paul, MN: Advanced Psychometrics.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice test by using error similarity analysis. *Teaching of Psychology, 16*(3), 151-155.

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. *Année psychol., 1905, 11*, 191-244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Chance News 4.01. (n.d.). Retrieved April 8, 2004, from *http://www.dartmouth.edu/~chance/chancenews/recentnews/chance_news_4.01.html#Computer%20admissions*

Chang, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398.

Chang, H. H., & Zhang, J. (2002, April). *Assessing CAT security breaches by the item pooling index –to compromise a CAT item bank, how many thieves are needed?* Paper presented at the meeting of the National Council of Measurement on Education, Chicago.

Cizek, G. J. (1999). *Cheating on tests: how to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Cohen, J. (1988). *Statistical power analysis for the behavioral science* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Publishers.

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T., Potenza, J. J. Fremer, & W. C. Ward. (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165–191). Mahwah, NJ: Lawrence Erlbaum Publishers.

Davey, T., & Stone, E. (2005, April). *A trend model for item parameter drift*. Paper presented at the meeting of the National Council of Measurement on Education, Montreal, Canada.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case study in computer adaptive test design through simulation* (Research Report 93-56). Princeton, NJ: Educational Testing Service.

Glas, C. A. W. (1998) Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8 (1),*. 647-667.

Glas, C. A. W. (1999). Item calibration and parameter drift. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 183-199). Boston: Kluwer Academic Publishers.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications*. Boston: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Han, N. (2003). *Using moving averages to assess test and item security in computer-based testing* (Center for Educational Assessment Research Report No. 468). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegation of copying* (Research Report Series No. 87-15). Iowa City, IA: American College Testing Program.

Holland, P. W. (1996). Assessing unusual agreement between incorrect answers of two examinees using K-index: statistical theory and empirical support (ETS Technical Report No. 96-4). Princeton, NJ: Educational Testing Service.

Isham, S. P., & Donoghua, J. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*, 33-51.

Levine, M. V., & Rubin, D. B. (1979). Optimal appropriateness measurement. *Psychometrika, 53*, 161-176.

Linn, R. L., Drasgow, F., Camara, W., Crocker, L., Hambleton, R. K., Plake, B. S., Stout, W., & van der Linden, W.J. (2002). CBT: A research agenda. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman, (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139-183). New York: Harper and Row.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Lu, Y., & Hambleton, R. K. (in press). Statistics for detecting disclosed items in a CAT environment. *Metodologia de Las Ciencias del Comportamiento*.

Meijer, R. R., & Sijtsma, K. (1995). Detecting of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, 8, 261-272.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz statistic to person-fit measurement. *Applied Measurement in Education, 21*, 115-127.

Paper-based GRE general test returning to parts of Asia. (n.d.). Retrieved October 15, 2002, from *http://www.ets.org/news/grecbt.html*.

Pitoniak, M. (2002). *Automatic item generation methodology in theory and practice* (Center for Educational Assessment Research Report No. 444). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Rupp, A. A., & Zumbo, B. D. (2003a, April). *Bias coefficients for lack of invariance in unidimensional IRT models*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Rupp, A. A., & Zumbo, B. D. (2003b). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *The Alberta Journal of Educational Research, 49*, 264-276.

Schnipke, D. L., & Scrams, D. J. (1999). *Item theft in a continuous testing environment: what is the extent of the danger?* (Law School Admission Council Computerized Testing Report). Newtown, PA: Law School Admission Council.

Segall, D. O. (2001, April). *Measuring test compromise in high-stakes computerized adaptive testing: a Bayesian strategy for surrogate test-taker detection.* Paper presented at the meeting of the National Council on Measurement in Education, Seattle, WA.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, 27, 163-179.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*(4), 439-460.

Smith, R. L., Wang, M. M., Wingersky, M., & Zhao, C. (2001, April). *Monitoring items for changes in performance in computerized adaptive tests*. Paper presented at the meeting of the National Council on Measurement in Education, Seattle, Washington.

Snijders, T. A. B. (in press). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*.

Sotaridona, L. S., & Meijier, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*, 53-69.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing pools* (ETS Research Report No 94-5). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23, 57-75.

Stocking, M. L, Ward, W. C., & Potenza, M. T. (1998). Simulating the use of disclosed items in computerized adaptive testing. *Journal of Education Measurement, 26*, 48-68.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston: Kluwer Academic Publishers.

Veerkamp, W. J. J. (1996). Statistical methods for computerized adaptive testing, Unpublished doctoral thesis, Twente University, the Netherlands.

Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25*, 373-390.

Wainer, H. (2000). Rescuing computerized adaptive testing by breaking zip's law. *Journal of Educational and Behavioral Statistics, 25*, 203-224.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R. J., et al. (Eds.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Publishers.

Wang, M. M., Wingersky, M., Steffen, M., & Zhu, R. (1998). *Preliminary operational item monitoring procedures.* Unpublished manuscript.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practices, Winter, 17-27.*

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Wells, C. S, Subkovik, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77-87.

Wolleck, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307-320.

Yi, Q., Zhang, J., & Chang, H. H. (2005, April). Identifying practical indices for enhancing item pool security. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

Yi, Q., & Chang, H. H. (2003). A-stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, 56, 359-378.

Zenisky, A.L., & Hambleton, R. K. (2004, April). *Investigating the effects of selected multistage test design alternatives on credentialing outcomes*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Zhu, R., Yu, F., & Liu, S. (2002, April). *Statistical indexes for monitoring item behavior under computer adaptive testing environment.* Paper presented at the meeting of the American Educational Research Association, New Orleans.

